vehicles

MDPI

Article

Radar–Camera Trajectory Matching for Roadside Units Using Motion Similarity and Self-Scoring Geometric Consensus

Zeynab Rokhi [1,*] and Ali Emadi [2]

[1] Department of Mechanical Engineering and McMaster Automotive Resource Centre (MARC), McMaster University, Hamilton, ON L8S 4L8, Canada
[2] Department of Electrical and Computer Engineering and McMaster Automotive Resource Centre (MARC), McMaster University, Hamilton, ON L8S 4L8, Canada; emadi@mcmaster.ca
* Correspondence: rokhiz@mcmaster.ca

Abstract

Roadside units that combine a millimetre-wave radar with a camera are becoming part of the road infrastructure, and using the two sensors together requires their extrinsic calibration. Above live lanes a target-based calibration is costly and has to be repeated, so calibration from the observed traffic, without targets, is the practical alternative. Any such method must first decide which radar track and which camera track belong to the same vehicle. Some published methods score each candidate pair from a pose fitted to that pair, which can be weakly constrained by a nearly straight vehicle track. All vehicles move on one road plane, so one plane-to-image homography must explain every correct match, and this paper builds the correspondence on that constraint. Calibration-free similarity cues rank the candidate pairs, the best-ranked pairs induce homography hypotheses, and each hypothesis is scored, without reference data, by the total cost of a global assignment of all radar tracks to all camera tracks. On eight hand-labelled recordings from one rooftop unit, the method assigns 656 of 672 radar tracks to the correct vehicle and selects the most accurate hypothesis on seven of them.

Keywords: data association; homography; intelligent transportation systems; radar–camera; roadside infrastructure; targetless calibration; track association

1. Introduction

Fixed sensing units that observe a road segment continuously are becoming part of the infrastructure of intelligent transportation systems. They feed traffic-adaptive signal control, incident detection, and cooperative perception services that extend the view of connected vehicles beyond their own sensors [1–4], as well as trajectory prediction at cooperative intersections [5]. A millimetre-wave radar next to a monocular camera is a common sensor pair for such units. The radar measures range and radial velocity in all weather, the camera provides the appearance needed for detection and classification, and the pair costs a fraction of a lidar installation [6–8]. Using the two sensors together requires their measurements to be expressed in one coordinate frame and therefore the extrinsic calibration between them. For a unit mounted several metres above live lanes, a target-based calibration means lane closures and trained staff, and the alignment drifts with temperature, wind, and maintenance, so the calibration has to be repeated [9,10]. Calibration from the traffic the unit already observes, without targets, is the practical alternative, and the recent literature on radar–camera calibration for roadside infrastructure, reviewed in [11], has moved in that direction.

Check for updates

Academic Editor: Hong Zhang

Received: 17 October 2026
Revised: 11 November 2026
Accepted: 13 November 2026
Published: 15 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Every targetless method of this kind rests on a decision that has to be taken before any calibration exists: which radar track and which camera track describe the same vehicle. The two sensors report incommensurable quantities, metric positions on a scan plane on one side and pixel coordinates on the other, and without the calibration there is no direct way to compare them. Most published track-to-track methods decide each candidate pair on its own, from a pose or a descriptor fitted to that pair alone. The most developed of them fits a camera pose to the candidate pair by the perspective-n-point (PnP) method and accepts the pair if that pose also explains a validation track [12]; others gate pairs by the azimuth of the target and its rate of change [13] or match discrete trajectory events such as lane changes [14]. These procedures were developed and validated on vehicle-mounted platforms, on pedestrians close to the sensors, or on multi-lane roads with frequent lane changes, where an individual trajectory carries enough depth and lateral structure to constrain a pose or to produce distinctive events. Shi et al. [15] note that trajectory-based methods "typically struggle in busy traffic scenarios with multiple unknown targets, as establishing accurate correspondences for calibration without prior knowledge remains a significant challenge".

The recordings studied here contain predominantly lane-keeping traffic viewed from an elevated roadside unit. Traffic is confined to lanes on a common surface, most vehicles keep their lane, and the elevated oblique viewpoint compresses the far end of the scene into a few pixel rows. An individual track is therefore nearly a straight line in both sensors, and a straight line does not determine a six-degree-of-freedom pose. Section 3 develops the geometric argument. Section 7 evaluates the proposed correspondence, on the eight labelled recordings and on a further recording from another day, and compares the prior information required by published calibration methods, together with a limited measurement of the association steps of two of them, reimplemented from their papers.

The same geometry supports a much stronger test. All vehicles move on the road plane, and a plane maps to the image by a single homography, which planar data determine exactly. One homography must therefore explain every correct match in a recording at once. The method of this paper is built on that observation. Similarity cues that need no knowledge of the calibration rank the candidate pairs. The best-ranked pairs, taken two at a time, induce hypotheses of the road-plane-to-image homography. Under each hypothesis every radar track is projected into the image and a gated global assignment of all radar tracks to all camera tracks is solved. The hypothesis with the lowest total assignment cost is selected, refined on all its matches, and completed by attaching the camera tracklets that occlusion had fragmented. The cost is computed from the data alone, and on every recording it tracks the accuracy of the matching it produces, so the procedure selects its own solution without reference data.

Figure 1 gives an overview of the correspondence procedure, from the separate sensor tracks to matched radar–camera trajectories.

The contributions of this paper are the following.

- A correspondence procedure for lane-bound roadside traffic that replaces the per-pair decision by a global one, a consensus over road-plane homography hypotheses scored by the total cost of a gated assignment, together with the calibration-free cues that propose the hypotheses.
- A manually labelled correspondence reference for eight recordings from an operational rooftop unit in three camera presets, 672 radar tracks with a camera counterpart, made against the synchronized video, with the labelling protocol and its limitations stated, and an independent blind relabelling of one recording that measures the agreement of the labels.

- An evaluation of what the selection criterion can and cannot do: the relation between the assignment cost and the accuracy of every hypothesis, the rank of the selected hypothesis, the sensitivity of the result to every parameter of the consensus stage, the contribution of each cue, and a failure analysis of the tracks that are missed or mis-assigned.
- A comparison of the information requirements and assumptions of three published roadside calibration methods [12–14] and the proposed correspondence stage, with a limited, caveated measurement of the association steps of the two track-level methods on the same recordings.

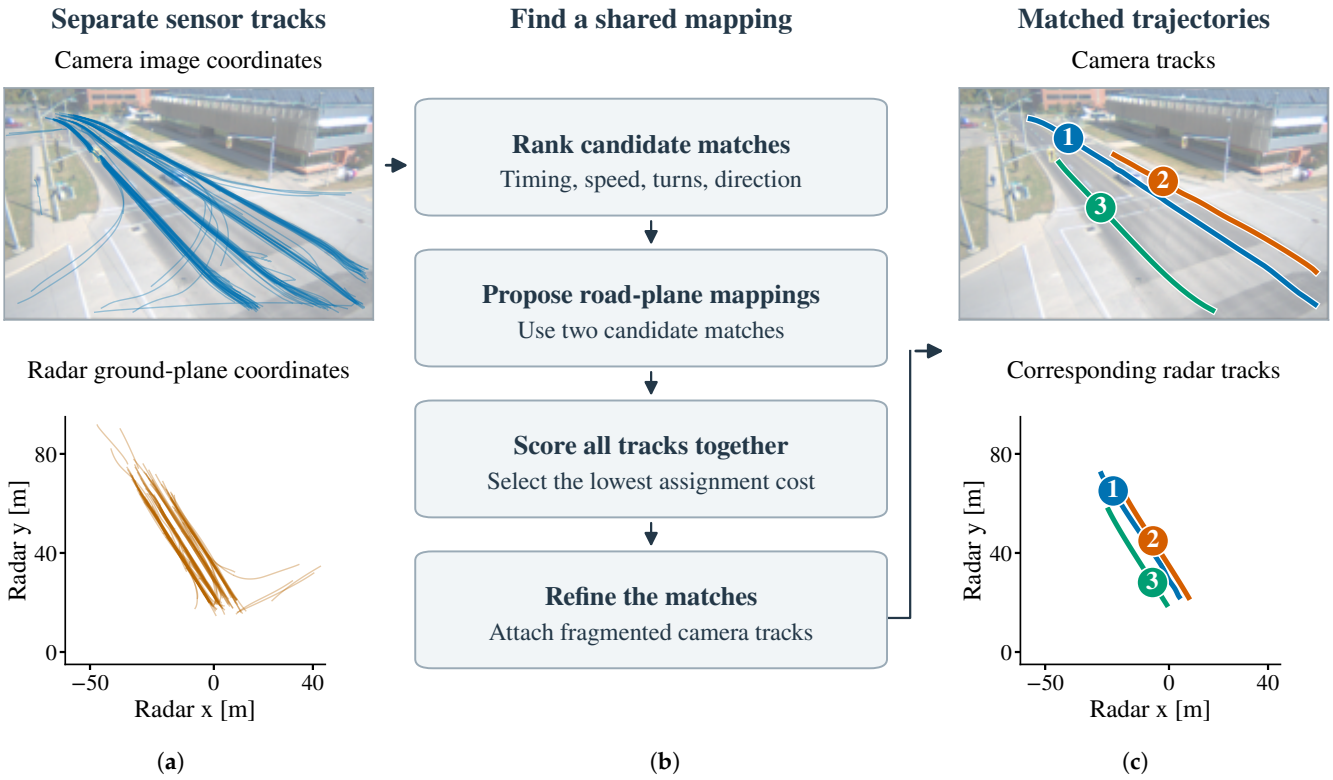


Figure 1. Overview of trajectory matching on recording R2. (**a**) Separate camera and radar tracks with unknown cross-sensor identities. (**b**) Motion cues propose candidate matches; a shared road-plane mapping tests them through global assignment, followed by refinement and tracklet attachment. (**c**) Three correctly matched output pairs, selected for illustration; colour and number identify the same vehicle across sensors. Trajectories span the recording and are drawn over a single background frame. The mapping supports correspondence, rather than providing the full sensor calibration.

The correspondence procedure is the first stage of a calibration from traffic, which then estimates the rigid transform, the camera focal length, and the radar–camera time offset from the matched pairs; that estimation is outside the scope of this paper. The present paper is concerned with the correspondence alone, which has its own reference data, its own failure modes, and its own comparison with published work. Section 2 reviews track-to-track association for radar–camera calibration. Section 3 states the problem and the geometric argument. Sections 4 and 5 describe the cues and the consensus. Section 6 describes the unit, the recordings, the labelling, and the scope of the literature comparison; Section 7 the results; and Section 8 the limitations.

2. Related Work

2.1. Correspondence in Radar–Camera Extrinsic Calibration

Extrinsic calibration between a radar and a camera is a two-stage problem in most of the literature: a correspondence stage that produces pairs of measurements of the same physical point or object in both sensors, and an estimation stage that fits the rigid transform to them [11]. Target-based methods obtain the correspondences from purpose-built artefacts [16–18]; targetless methods take them from objects already in the scene, first from instantaneous detections of moving objects [19,20] and later from whole tracks, which carry the temporal context that separates vehicles close in space [14,21]. The review in [11] found that few of these studies report the accuracy of the correspondence step itself, and that most treat the quality of the matching as given. The correspondence is the quantity evaluated in this paper.

2.2. Track-to-Track Association Without a Calibration

Classical track-to-track association was developed for distributed tracking with a known frame alignment [22]; the calibration setting removes that assumption. The published approaches for radar–camera pairs fall into three groups.

Geometry-scoring methods evaluate a candidate pairing by how well a calibration solution computed from the pair explains the data. Liu et al. [12] solve a PnP problem for every candidate radar–camera track pair and score the pair by the reprojection consistency of the resulting pose on the pair itself and on a validation track. The method needs no prior extrinsics, reaches an association accuracy of 96% on pedestrian sequences, and was validated for calibration on the vehicle-borne CARRADA data set [23]; its estimator was later reused for pedestrians with the time offset obtained from a cooperative gesture [24]. Cao et al. [13] form the first correspondences from the azimuth angle of each target and its rate of change, which presumes that the radar boresight and the camera optical axis lie in one vertical plane, then refine the matching by a Hungarian assignment in the image once a coarse ground-plane PnP solution exists; the camera intrinsics and the installation height are given and the lane heading is taken from the traffic. Jin et al. [25] switch between a minimum-distance matching at low traffic volume and a Hungarian assignment at high volume, with the camera side localized in three dimensions through annotated vehicle models. In all three the quality of the association depends on the conditioning of a pose fitted to one track or to a few frames.

Event- and shape-matching methods compare trajectory descriptors instead. Zhu et al. [14] detect a vehicle crossing a lane line in both sensors, recover the distance of each camera event along the road by the cross ratio from the vanishing point and the known length of the lane markings, and match events in a frame–distance domain; the method needs the lane lines in the image, the number of lanes, and hardware synchronization, and it depends on lane changes. Du et al. [26] generate events at virtual detection lines drawn on the lane geometry and match vehicles by the headway to their neighbours, which also recovers the time offset, but the output is a plane mapping rather than a six-degree-of-freedom extrinsic. Shape descriptors such as curvature are not preserved under perspective projection, and the distortion grows with the distance from the optical axis, which for an elevated roadside camera is where most of the traffic is seen. Durmaz and Cevikalp [27], who align the trajectory of one cooperative drone observed with a depth camera, note that straight motion makes the alignment degenerate. Shi et al. [15] match Kalman-smoothed trajectories on the ground plane by motion-state consistency, with the camera trajectories obtained by inverse perspective mapping, so a camera-to-ground homography must already be known; no calibration accuracy is reported. FusionCalib [21] reconstructs the road plane from associated vehicle bottoms but projects the radar into the

image with a given radar–camera transform, so it calibrates the sensors with respect to the road rather than to each other.

Graph- and learning-based methods represent the traffic scene as a network. Huang et al. [28] form directed graphs of radar and camera trajectories weighted by radar cross-section, Zhao et al. [29] match the roadside detections by graph matching to the perception and high-precision positioning of a cooperative autonomous vehicle, which supplies the reference that a roadside unit alone lacks, and Cheng et al. [30,31] train a discriminator that decides whether a radar and a camera detection belong to the same object. A recent roadside formulation avoids explicit matching by aligning Gaussian models of the targets with a distribution distance [32]. Learned associations require training data representative of the deployment viewpoint. Public radar–camera data sets include vehicle-mounted recordings [23,33] and the roadside subset V2X-Radar-I [34], but their viewpoint and visible trajectory extent need not match those of the elevated unit studied here. Section 8 explains why the V2X-Radar-I sequences examined were not used for quantitative evaluation.

2.3. Relation to Prior Work

The procedure of this paper scores a correspondence hypothesis by the global assignment it induces under a shared geometric model. The pattern is an old one. Hypothesize-and-verify estimation, of which RANSAC is the standard form [35], fits a model to a minimal sample and scores it by the agreement of all data, and the graph-alignment methods above score a matching by a global objective. The present method combines the two: the minimal sample is a pair of candidate track matches, the model is the road-plane homography [36], and the score is the cost of a gated linear assignment of all tracks under that model, so the matching and the model are decided together. It never fits a pose to a single track, which is what the geometry-scoring methods do and what lane-bound tracks do not support. Compared with the event methods, it uses every vehicle that both sensors see rather than the few that change lane, and it needs no lane lines, no marking lengths, and no hardware synchronization. Section 7.9 compares the information requirements of these published methods with those of the proposed correspondence stage, including the temporal alignment and common-plane assumptions retained here.

3. Problem Statement and Geometric Model

3.1. Setting and Notation

A radar and a camera are mounted together above a road and observe the same traffic. Each sensor is processed independently up to the level of object trajectories (Section 4). The input to the correspondence stage is a set of radar trajectories $\{\mathcal{T}_i^r\}_{i=1}^{N_r}$, each a sequence of samples $(\mathbf{p}_k, \dot{r}_k, t_k)$ with a planar position $\mathbf{p}_k = (x_k, y_k)$ in the radar frame, a Doppler velocity $\dot{r}_k$ and a time stamp t_k, and a set of camera trajectories $\{\mathcal{T}_j^c\}_{j=1}^{N_c}$, each a sequence of samples $(\mathbf{q}_k, w_k, t_k)$ with the bounding-box centre $\mathbf{q}_k$ in pixels and the box width w_k. Both sets are indexed by the synchronized frame number so that two tracks active in the same frame refer to the same instant up to the radar–camera time offset, which is small compared with the duration of a track. The task is to find the set of pairs $\mathcal{M} \subset \{(i, j)\}$ that describe the same vehicle. Nothing is known about the relative pose of the two sensors, the camera focal length is known only nominally, and no manual initialization, rectification or label is used.

Camera tracks fragment. A vehicle that is occluded or briefly missed by the detector reappears with a new identifier, so one radar track may legitimately correspond to several camera tracklets, and N_c is roughly one and a half to two times N_r on the recordings of Section 6. The method has to accommodate this.

3.2. Road-Plane Homography as the Geometric Model for Correspondence

The published method closest to this setting decides each correspondence on its own [12]: a camera pose is computed from the candidate pair by PnP, and the pair is accepted if that pose also explains a validation track. Consider what a single lane-bound track constrains. Its radar samples lie, to within the lateral scatter of the return, on a line segment in the road plane, and its image samples lie on a line segment in the image. The set of camera poses that map one segment onto the other is far from a point. A rotation about the line through the radar samples leaves every one of them in place, so the reprojection Jacobian with respect to the pose is rank deficient and a one-parameter family of poses fits the pair exactly. The directions that move the far end of the segment in depth are almost as weakly constrained, because the oblique view compresses that end into a few pixel rows. The PnP solver returns one member of the family, selected by the noise of the samples. Applied to the validation track, the members of the family project it to very different places, so the validation error of a true pair is spread over a wide range, and the same holds for a wrong pair whose own points were fitted just as well. This is a property of the geometry rather than of an implementation, and it is the reason the method of this paper never fits a pose to a single track.

What the roadside scene offers instead is a constraint that is weak for one track and strong for all tracks together. The vehicles move on the road surface, which is a plane to good approximation over the area a unit observes, and a plane in the world maps to the image plane by a homography $\mathbf{H} \in \mathbb{R}^{3\times3}$ [36]: for a radar position $\mathbf{p} = (x, y)$ on the road plane and its image $\mathbf{q}$,

$$\begin{bmatrix} \mathbf{q} \\ 1 \end{bmatrix} \sim \mathbf{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{1}$$

where $\sim$ denotes equality up to scale. The homography has the properties the correspondence problem needs. It is global: a single $\mathbf{H}$ has to explain every matched pair in the recording, and a wrong hypothesis that happens to fit one pair projects all the other radar tracks onto the wrong lanes. It is determined by planar data: it has eight degrees of freedom and is fixed by four point correspondences in general position on the plane, whereas the same data leave the uncalibrated 3×4 resection rank deficient and a calibrated pose from a single nearly straight track poorly constrained. It can be formed from two matched tracks with non-collinear combined extent, each contributing tens of samples along its path, so that a hypothesis arises from a pair of candidate matches and can be tested against the whole scene. The homography is not the calibration itself, which requires the camera intrinsics and treats the height of the vehicle bodies above the plane and the time offset properly; it is the coarse geometric model against which correspondence hypotheses are judged.

The plane in question is, strictly, not the road surface but the plane of the reference points the two sensors supply: the radar reports the scattering centre of the vehicle body and the camera the centre of the bounding box, both of them above the road at roughly half the height of the vehicle. A plane parallel to the road is still a plane, and its points reach the radar's planar coordinates by a parallel projection, which is affine, so the homography fitted to the data is that of the elevated plane and a reference height common to all vehicles is absorbed exactly. What is not absorbed is the variation of that height from vehicle to vehicle, a car against a truck, and the offset between the scattering centre and the box centre along the body. For a camera about 10.8 m above the road, half a metre of height moves the image point by about 10 pixels at 30 m and 5 pixels at 60 m, almost entirely along the image vertical. Section 5 states how the assignment cost tolerates shifts of this size, and Section 7.4 measures them.

The correspondence problem is therefore posed as follows: generate a small set of candidate homographies from pairs of plausible track matches, score each candidate by how well one model explains all tracks at once, and select the candidate with the best score. The next two sections describe the cues that propose the candidates and the consensus that decides among them.

4. Trajectories and Calibration-Free Similarity Cues

4.1. Per-Sensor Trajectories

Camera frames and radar scans are paired by nearest time stamp within a 100 ms gate, and the true time stamp of every frame is retained, because the measured frame interval of the camera jitters by about a third of its mean on the unit of Section 6. The camera is triggered by software on the arrival of each radar frame, so the paired time stamps agree to about a millisecond and the offset that remains between the two streams is the latency of the radar's own processing; it is not measured independently for these recordings, and the calibration stage that follows the correspondence places it at one to two frame intervals; Section 7.5 tests how far the correspondence tolerates a shift of the camera stream by whole frames. Radar detections are restricted to moving points, i.e., points with non-zero Doppler, which removes the static scene without any spatial mask. The moving points of each scan are clustered with DBSCAN [37], and the cluster centroids are tracked by a global-nearest-neighbour Kalman tracker whose assignment is gated on position and on Doppler velocity, so two vehicles that are close in space but differ in radial speed are not merged.

Camera frames are processed by a pretrained vehicle detector, YOLO26 [38], and the detections are linked by the SORT tracker [39]; each camera track stores the bounding box and its centre at every frame. Both track sets are smoothed with arc-length splines followed by a conservative curvature moving average [40], so positions, speeds, and headings are available as continuous functions of time, and stationary segments are masked, in the radar branch from the Doppler velocity and in the camera branch from the pixel velocity. The detector and the trackers are standard components; the correspondence method assumes only a bounding box per detection and a time stamp per sample.

4.2. Similarity Cues

Every radar–camera track pair is first scored with cues that can be evaluated without knowing where the radar frame lies in the image. They are approximately viewpoint-robust rather than exactly view-independent: the normalizations below reduce, but do not remove, the dependence on viewing direction and depth, and the cue ablation of Section 7.7 measures their contribution. They are cheap to compute and not reliable enough to decide a correspondence on their own; their role is to rank candidates for the consensus stage. Four cues are computed on the frames the two tracks have in common.

Time overlap. With I_i^r and I_j^c the frame intervals over which the two tracks are active, the time score is the intersection over union of the intervals,

$$s_t(i,j) = \frac{|I_i^r \cap I_j^c|}{|I_i^r \cup I_j^c|}, \tag{2}$$

which is zero for tracks that never coexist and removes most of the $N_r \times N_c$ candidate space in a recording of a few minutes. It is the one cue that is exact, up to the time offset between the sensors and the fragmentation of the camera tracks, and it receives the largest weight.

Speed profile. Speed is comparable across the two sensors only after the unknown depth is removed from the camera side. The pixel speed of a bounding-box centre divided by the

box width is, to first order, the ground speed divided by the physical width of the vehicle, because both the displacement and the box scale with the inverse of the depth:

$$\frac{v_{\mathrm{px}}}{w} \approx \frac{v}{W}, \tag{3}$$

with W the vehicle width. The ratio removes the leading dependence on depth, so no image-to-ground rectification is needed; with a nominal width of 1.8 m it is expressed in the units of the radar speed. The proportionality holds only up to a factor that depends on the viewing direction, because the pixel displacement of a point moving along the line of sight falls with the square of the depth whereas the box width falls with the depth. For the geometry of the unit of Section 6, the factor is about 0.4 at every depth for crossing traffic and falls from 0.6 at 10 m to 0.13 at 80 m for a vehicle approaching along the road, and on the labelled pairs the measured ratio of the normalized camera speed to the radar speed has a median of 0.28, from 0.33 at 20 to 32 m to 0.21 at 47 to 75 m (Figure 2). The nominal width therefore acts as a scale parameter of the cue rather than as a physical quantity; Section 7.3 shows that the correspondence is insensitive to it. Both speeds are then quantized into four states (stationary, slow, moderate, fast, with thresholds at 0.1, 0.4 and 1.5 m of displacement per mean frame interval), and the score is

$$s_v(i,j) = 1 - \frac{1}{3K}\sum_{k=1}^{K} |b_k^r - b_k^c|, \tag{4}$$

where b_k^r and b_k^c are the state indices of the two tracks at common frame k and K is the number of common frames. The quantization makes the score insensitive to the exact vehicle width and to the noise of a differentiated pixel trajectory while keeping the information that matters, whether the vehicle is stopped, creeping, or moving freely at a given time. With the factor of about 0.3, the camera states lie one state below the radar states for most moving vehicles (mean state 1.1 against 2.0 over the common frames of the labelled pairs), so in practice the cue separates stopped and creeping vehicles from moving ones rather than slow from fast; the mean state difference is 0.89 to 0.94 for the labelled pairs against 0.80 to 1.03 for all other overlapping pairs on the eight recordings. This is why the cue ranks weakly on its own and why the decision is left to the geometry; eight states or a continuous score do not do better (Section 7.3).

Turning events. Curvature magnitude is not preserved by perspective projection: the same physical turn produces a large pixel curvature near the camera and almost none at depth or at the image periphery, so a comparison of curvature profiles, used in an earlier version of this work, was unreliable exactly where roadside cameras see most of their traffic. What projection preserves is the event. A straight line stays a straight line, and the moment at which a vehicle begins and ends a turn, and the direction of the turn, survive projection. Each track is therefore annotated per frame with a turn state in $\{-1, 0, +1\}$ (turning one way, straight, turning the other way) from the median-filtered heading-change rate of the radar trajectory and of the image trajectory, respectively, with a threshold of 0.06 rad/s and a minimum speed below which the heading is undefined. The score $s_e(i,j)$ is the per-frame agreement of the two state sequences over the common window: credit for two tracks turning the same way at the same time, a neutral value of 0.5 when fewer than eight frames carry a defined state on both sides, and a penalty for opposite turns. The image left–right sign may be mirrored with respect to the radar axes depending on the mounting; it is estimated once per recording from the turn-rate correlation over the candidate pairs rather than assumed.

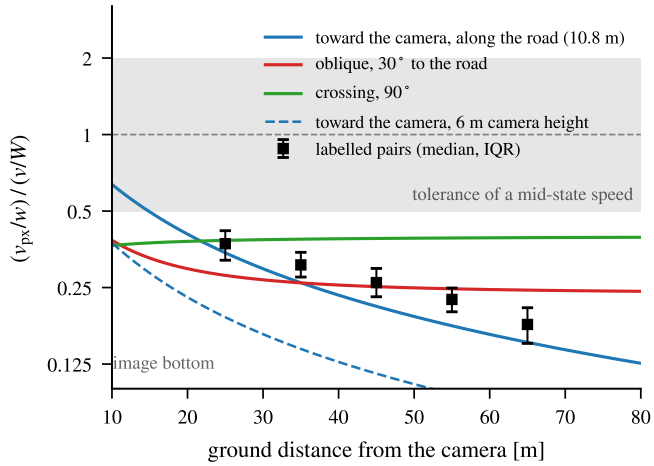


Figure 2. Validity range of Equation (3). Lines: the factor $(v_{px}/w)/(v/W)$, which the equation takes as one, for a pinhole camera 10.8 m above the road (the rooftop of Section 6 and its mount) pitched 23° down with $f = 667$ pixels and a vehicle of $1.8 \times 4.5 \times 1.5$ m, against the ground distance, for three headings; dashed, the same camera mounted at 6 m. Squares: the measured ratio of the normalized camera speed to the radar speed over the 832 labelled pairs of the eight recordings that contain moving frames, median and interquartile range per 10 m of range. Shaded band: the factor within which a speed in the middle of a quantization state stays in its state, since consecutive thresholds differ by a factor of about four.

Direction of travel. A radar track whose range decreases belongs to an approaching vehicle, and an approaching vehicle grows in the image. The radar range rate $\mathrm{d}\ln r/\mathrm{d}t$ and the negative logarithmic rate of change of the box width, $-\mathrm{d}\ln w/\mathrm{d}t$, are therefore compared as a graded score $s_d(i,j) \in [0,1]$, with 0.5 when either side is uninformative, and blended toward 0.5 in proportion to the median range rate so that crossing traffic, whose range barely changes, does not receive a spurious verdict. The same comparison is applied as a hard veto: a pair in which the radar track approaches while the box shrinks, or recedes while the box grows, is excluded before any further scoring, provided both sides are individually informative (at least ten common frames, a range rate above 0.8 m/s with a consistent sign in at least 70% of the samples, and a relative box-width change of at least 15% with a consistent sign in at least 65% of the samples). Figure 3 shows the four cues for a true pair and for a wrong pair in the same lane.

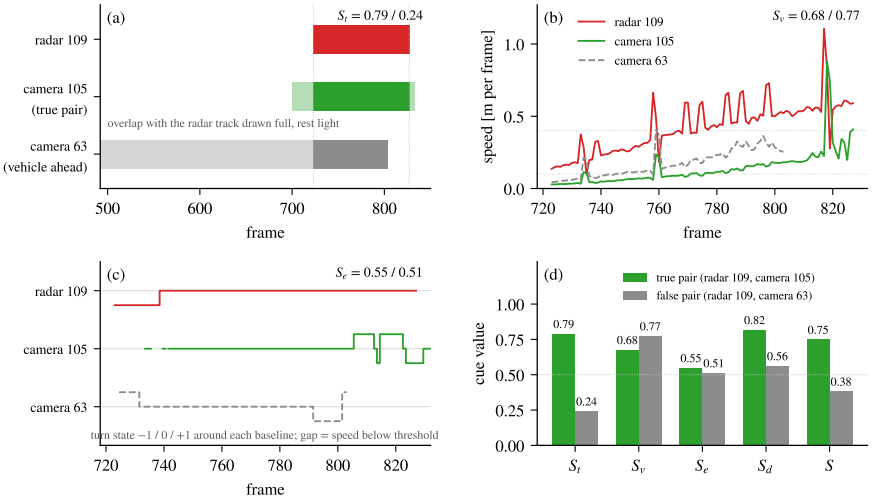


Figure 3. The four cues for one radar track of recording R2 (track 109) against its labelled camera track (105) and against the camera track of the vehicle 13 m ahead in the same lane (63). (**a**) Frame intervals

of the three tracks; the part of each camera track that overlaps the radar track is drawn full. (**b**) Speed over the common frames, the radar speed and the camera pixel speed normalized by the box width, in the units of the speed cue; the dotted lines are the quantization thresholds. (**c**) Turn state per frame (-1, 0, $+1$ around each baseline; gaps where the speed is below the threshold). (**d**) The resulting cue values and the combined similarity S. For a same-lane neighbour, the speed and turning cues do not separate the two candidates; the time overlap and the direction score do, and the geometry of Section 5 decides the rest.

4.3. Combined Similarity Score

The combined similarity is a weighted mean,

$$S(i,j) = \frac{5\,s_t + s_v + s_e + s_d}{8}, \tag{5}$$

with the time overlap weighted most heavily because it is the one cue that is exact. Pairs with $S < 0.1$ and vetoed pairs are removed from consideration. Similarity alone is a weak decision rule. On the eight labelled recordings of Section 6, a labelled pair scores above $S = 0.5$ for only 53 to 65% of the radar tracks (Section 7.1), so a threshold on the cues would leave a third of the tracks unresolved, and among the fifteen highest-scoring pairs of a recording, only four to six are true pairs (Section 7.2). Vehicles in adjacent lanes that enter and leave the field of view together have the same timing, the same speed profile, no turns, and the same direction, and no calibration-free cue can tell them apart. The ranking is nevertheless sufficient for the next stage, which needs only two true pairs among the seeds to form a correct hypothesis and then decides the matching for every track by geometry.

5. Consensus over Road-Plane Homography Hypotheses

Figure 1 summarizes the stage. The cues of Section 4 produce a seed set, pairs of seeds induce homography hypotheses, each hypothesis is scored by the total cost of a global assignment, and the hypothesis with the lowest cost is selected, refined on all its matches, and completed by attaching unassigned camera tracklets. Every step is described below with the values of its parameters; Section 7.3 reports how much the result depends on each of them.

5.1. Seed Set and Hypothesis Generation

The $K_s = 15$ candidate pairs with the highest similarity, among those with $S \geq S_{\min} = 0.5$ and at least 30 common frames, form the seed set $\mathcal{S}$. Every combination of two seed pairs that involves two distinct radar tracks and two distinct camera tracks defines a hypothesis, provided the radar points of the two tracks together span the plane: the smaller singular value of their centred coordinates must exceed 3 m, so two tracks in the same lane, whose combined points are nearly collinear and would leave the homography undetermined across the road, are not used. For each admissible combination, a homography $\mathbf{H}$ is fitted to the stacked samples of the two pairs by the normalized direct linear transform inside a RANSAC loop with a 5 px inlier threshold [35,36]; the loop guards against the few samples of a seed that an occlusion or a detector jump has displaced. There are at most $\binom{15}{2} = 105$ combinations, 71 to 89 of them admissible on the recordings of this paper, and each is evaluated in a few milliseconds. A hypothesis is correct only if both of its seeds are true pairs, so with four to six true pairs among fifteen seeds, three to thirteen of the hypotheses are correct, and the selection has to find one of them among the rest.

5.2. Global Assignment Under a Hypothesis

Under a hypothesis $\mathbf{H}$ every radar track is projected into the image, and the distance between a radar track and a camera track is measured there:

$$d_{\mathbf{H}}(i,j) = \frac{1}{K_{ij}} \sum_{k \in I_i^r \cap I_j^c} \left\| \mathbf{q}_{j,k} - \pi(\mathbf{H}\,\tilde{\mathbf{p}}_{i,k}) \right\|, \tag{6}$$

the mean Euclidean distance in pixels over the K_{ij} common frames between the camera bounding-box centres and the projected radar positions, where $\tilde{\mathbf{p}}$ denotes homogeneous coordinates and π the dehomogenization. The distance enters the assignment through a Geman–McClure function [41] with scale $\sigma = 50$ px,

$$\rho(d) = \frac{(d/\sigma)^2}{1 + (d/\sigma)^2}, \tag{7}$$

which saturates at one for gross mismatches, so a wrong hypothesis is not dominated by a few enormous distances and the cost of a hypothesis measures how many tracks it explains rather than how badly it fails on the others. The scale is also large against the systematic shifts of Section 3: a shift of 10 pixels common to all frames of a pair costs $\alpha\rho = 0.19$ under the weights below, against a gate of $2c_0 = 5$, so a vehicle whose reference points sit higher or lower than the rest is still matched to its own track, whereas under a correct hypothesis the other overlapping pairs lie about 200 pixels away in the median (Section 7.2). The assignment cost of a pair combines geometry and similarity,

$$c_{\mathbf{H}}(i,j) = \alpha\,\rho\big(d_{\mathbf{H}}(i,j)\big) + \beta\,\big(1 - S(i,j)\big), \qquad \alpha = 5,\ \beta = 3. \tag{8}$$

A linear assignment between the radar and the camera tracks is then solved by the Hungarian method [42] on the $N_r \times N_c$ matrix of costs (8), augmented with a dummy column for every radar track and a dummy row for every camera track, each of cost $c_0 = 2.5$. Leaving a candidate pair unmatched therefore costs $2c_0$, so a pair is matched only when its cost is below $2c_0 = 5$, and a track whose best candidate costs more than that is left unmatched rather than forced onto a wrong partner. Pairs that were vetoed or that have no common frames receive a prohibitive cost. The result is an assignment $A(\mathbf{H})$ and its total cost

$$C(\mathbf{H}) = \sum_{(i,j) \in A(\mathbf{H})} c_{\mathbf{H}}(i,j) \;+\; c_0\, n_{\text{unmatched}}(\mathbf{H}), \tag{9}$$

where $n_{\text{unmatched}}$ counts the tracks of either sensor assigned to a dummy node.

5.3. Hypothesis Selection

The hypothesis with the smallest total cost is selected,

$$\mathbf{H}^\star = \arg\min_{\mathbf{H}} C(\mathbf{H}). \tag{10}$$

A wrong hypothesis places the projected radar tracks on the wrong lanes or off the road, so few camera tracks can be matched closely; most tracks are paid for at the gate cost or at the saturated geometric cost, and C is high. The correct hypothesis places every radar track on its own lane, most tracks match within a few pixels, and C is low (Figure 4). Section 7.2 shows that on every recording, the cost tracks the accuracy of the matching that each hypothesis produces, so the procedure selects its own solution. Whether the absolute level of the cost can also flag a recording on which no hypothesis is acceptable has not been tested, because the present data contain no such recording.

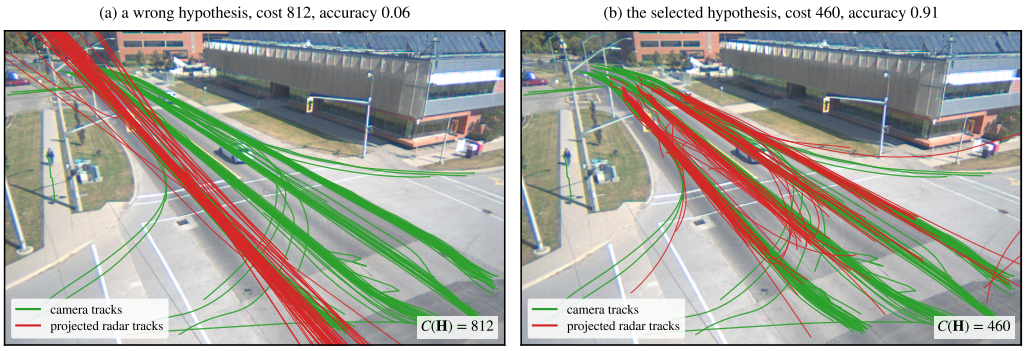


Figure 4. How the assignment cost identifies the correct correspondence (recording R2, 126 radar and 204 camera tracks). Camera trajectories and radar trajectories projected into the image through the homography of a seed hypothesis. (**a**) A wrong hypothesis: the projected radar tracks fall across the lanes, few camera tracks can be matched, and the total assignment cost is high. (**b**) The selected hypothesis: the projected radar tracks coincide with the camera tracks along the whole approach and the cost is the lowest of the 87 hypotheses. Both panels show a two-seed hypothesis before refinement, not the final matching (Section 7.8). The accuracies are known only because the recording is labelled; the costs are computed from the data alone.

5.4. Model Refinement and Tracklet Attachment

The selected hypothesis was fitted to two tracks. The assignment is therefore refined in two rounds: the homography is refitted by RANSAC on the samples of all matched pairs, the assignment is solved again under the refitted model, and the process is repeated once. Refinement changes the assignment of the few tracks whose projection under the two-seed homography was marginal, and on the recordings of this paper it changes the accuracy of the selected hypothesis by -1.0 to $+4.1$ points (Section 7.2). Finally, camera tracklets that remain unassigned are attached to an already matched radar track when their cost under $\mathbf{H}^\star$ is below a stricter gate $c_1 = 2.0$ and the pair is not vetoed. This is how the fragmentation of camera tracks is absorbed: a radar track may end up with several camera tracklets, each covering part of its duration, and all of them are passed on as separate pairs. The output of the stage is the set of matched pairs $\mathcal{M}$, the homography $\mathbf{H}^\star$ and the cost $C(\mathbf{H}^\star)$.

5.5. Computational Complexity and Run Time

The similarity stage evaluates the four cues on every pair of tracks with a common frame, of which there are 560 to 1320 on the recordings of this paper, and takes 0.8 to 3.3 s per recording in a single-threaded Python implementation. The consensus stage evaluates one homography fit, one projection of all radar tracks and one linear assignment of size $(N_r + N_c)^2$ per hypothesis, and takes 0.5 to 3.0 s per recording for 71 to 89 hypotheses. In order terms, the similarity stage costs $O(N_r N_c \bar{K})$ for $\bar{K}$ common frames per pair; the hypothesis generation at most $\binom{K_s}{2}$ homography fits on a few hundred points each, independent of the number of tracks; and every hypothesis involves one projection of all radar samples and one Hungarian assignment on the augmented $(N_r + N_c) \times (N_r + N_c)$ matrix, $O((N_r + N_c)^3)$ in the worst case, so the assignment dominates and the stage scales with the cube of the number of tracks. The refinement adds two assignments and the attachment one pass over the unassigned camera tracks. The whole correspondence stage therefore runs in a few seconds on a laptop for two to four minutes of traffic, against several minutes for the detection and tracking that precede it.

6. Experimental Setup and Reference Labels

6.1. Roadside Unit and Recordings

All data come from one roadside unit on the rooftop of the McMaster Automotive Resource Centre in Hamilton, Ontario, about 10.6 m above a signalized intersection and its approach. The unit combines a Retina-4FN imaging radar (Smart Radar Systems (Seongnam-si, Republic of Korea); 100° azimuth by 24° elevation field of view, point detections with Doppler velocity, 50 ms update period) and a Basler ace 2 a2A1920-51gcBAS camera (Sony IMX392 sensor (Sony, Tokyo, Japan), 1936 × 1216 native pixels) behind a Theia MY23F rectilinear lens, mounted together in one enclosure with the radar directly below the camera. The camera driver offers capture presets labelled by a nominal diagonal field of view; the presets are centre crops of the sensor at the native pixel scale, so their frames differ in size but not in focal length. The unit was never calibrated by a target-based procedure, no reference extrinsics exist, and the relative position of the two sensors was not measured.

Eight recordings of traffic were made on 3 October 2023 between 11:55 and 12:46, of 152 to 272 s each, in three presets and in the readout used for the reference recording (Table 1). They are identified as R0 to R7. R0 is the recording on which the method was developed and on which the earlier labels of this work were made, and R1 to R7 are the evaluation recordings. All eight carry manual labels made after the final version of the per-sensor trackers, so no label refers to a track that no longer exists. Each recording contains 57 to 126 radar tracks and 101 to 204 camera tracks; the camera side has about one and a half to two tracklets per radar track because occlusion and detector misses fragment the image tracks. A ninth recording, made on another day with the camera aimed differently, is reported apart from R0 to R7 in Section 7.1.

Table 1. The eight labelled recordings. Frame size in pixels; tracks after the per-sensor trackers of Section 4; every radar track was labelled. "None": radar tracks without a camera counterpart (vehicle outside the image, or a radar track that merges two vehicles); "two tracklets": radar tracks whose counterpart is fragmented into two camera tracklets; pairs: labelled radar–camera tracklet pairs.

				Tracks		Labelled Radar Tracks			
Rec.	**Preset**	**Frame [px]**	**Duration [s]**	**Radar**	**Camera**	**None**	**With Counterpart**	**Two Tracklets**	**Pairs**
R0	100°-class	1108 × 622	209	77	155	3	74	15	89
R1	90°	904 × 568	231	104	190	4	100	34	134
R2	90°	904 × 568	268	126	204	12	114	19	133
R3	100°	1076 × 676	201	82	141	2	80	13	93
R4	100°	1076 × 676	211	100	191	4	96	41	137
R5	100°	1076 × 676	152	57	101	2	55	14	69
R6	110°	1292 × 812	272	89	133	7	82	16	98
R7	110°	1292 × 812	178	73	108	2	71	10	81
All			1722	708	1223	36	672	162	834

6.2. Labelling Protocol

The initial labels were made by the first author against the synchronized video. For each recording a video was rendered in which every camera track is drawn with its identifier on the frames and every radar track is listed with its frame span, and a checklist was prepared with one row per radar track, giving its frame span, the camera track the automatic procedure had assigned to it, and every other camera track that overlaps it in time. The labeller watched the video and wrote, for each radar track, the identifier of the camera track that shows the same vehicle, two identifiers when the vehicle's image track is fragmented into two tracklets, or "no camera" when the vehicle is outside the image or is not detected or when the radar track itself merges two vehicles and has no single

counterpart. In total, 708 radar tracks were labelled, 672 of them with a camera counterpart, 162 of those with two tracklets, and 834 radar–camera tracklet pairs (Table 1).

The protocol has limitations that the reader should know. The checklist showed the automatic assignment, so the labeller confirmed or corrected it rather than labelling blind. To counter the bias this introduces, every disagreement between a label and the assignment was listed afterwards and re-examined a second time in the video, and the conflicts that remained were decided by the labeller without reference to the assignment. A camera tracklet that covers a vehicle for only a few frames at the image border is labelled only when its identity is unambiguous, so the recall figures below are measured against the tracklets a human could confirm. The labels are keyed by track identifier and refer to the final tracks; an earlier label set on the reference recording, made on the tracks of an earlier tracker version, was not transferred but relabelled.

To measure the bias directly, all 126 radar tracks of R2 were relabelled by a person who had taken no part in the work and had not seen the automatic assignments, from the same video but with a checklist that lists, for each radar track, its frame span and every camera track overlapping it in time, in identifier order, and nothing else. The independent labels agree with those of the paper on 123 of the 126 tracks (0.976) and on 124 when a radar track with two tracklets counts as agreed if the two labellers share at least one of them; on the decision whether a radar track has a camera counterpart at all, the two label sets agree on 124 tracks, Cohen's $\kappa = 0.91$. The three disagreements are one track that the independent labeller leaves without a counterpart and the paper's labels match, one that the independent labeller matches and the paper's labels leave without a counterpart, and one for which the independent labeller names one of the two tracklets of the paper's label. Scored against the independent labels instead of the paper's, the accuracy of the method on R2 is 0.947 (0.956 against the paper's labels) and its pair precision 0.935 (0.951). On the three tracks where the label sets differ, the paper's label coincides with the output of the method, the direction in which a confirmation bias would act; its effect on the accuracy of this recording is one radar track.

6.3. Metrics

Association quality is reported at two levels. Accuracy is counted per radar track: a radar track with a counterpart is correct if at least one of the camera tracklets assigned to it is a labelled counterpart, and accuracy is the fraction of correct tracks among those with a counterpart. Precision and recall are counted per radar–camera tracklet pair, which is the unit the estimation stage consumes: precision is the fraction of assigned pairs that are labelled pairs, and recall is the fraction of labelled pairs that are assigned. A radar track without a counterpart that receives a camera track counts against precision but not against accuracy. For the consensus stage two further quantities are measured on every recording: the accuracy of the matching that each seed hypothesis produces before refinement, which requires the labels, and the Pearson correlation r (with the Spearman rank correlation as a check) between the assignment cost $C(\mathbf{H})$ of the hypotheses and that accuracy, which tells whether the selection criterion is informative. Precision matters more than recall for the downstream estimation, because a wrong pair biases the calibration whereas a missed pair only reduces the amount of data.

6.4. Scope of the Literature Comparison

The literature comparison concerns the information and assumptions needed to apply the published methods. It separates external inputs from quantities obtained automatically from observations. In addition, the association steps of the two track-level methods were reimplemented from their papers, since the authors' code is not available, and run on the

eight recordings; Section 7.9 reports the result. In the reimplementation of Liu et al. [12], each candidate pair supplies a PnP pose, the score of the pair is the sum of its mean squared reprojection error and that of a validation track under the same pose, as in their Equations (10) to (14), each camera track is assigned to the radar track with the lowest score, and a pair is rejected when its validation error exceeds 40 pixels RMS; the validation tracks are the two longest camera tracks that move in at least 90% of their samples, a choice the paper leaves open, and the focal length is the native 666.7 pixels. In the reimplementation of Zhu et al. [14], radar trajectories are clustered into lanes by their Hough line features, lane changes are detected as segments between lane neighbourhoods on the radar side and as boundary crossings on the camera side, and events are matched greedily by a cost in frames and along-road distance; the lane lines and longitudinal distances are supplied from the surveyed geometry of the site, a stronger prior than the image segmentation of the original method. Each method is run in two representations: "published" uses the bottom centre of the bounding box and a zero time offset, "adapted" the box centre and the time offset that the calibration stage later estimates, and, for Liu et al., the fitted radar return plane instead of a constant radar height. The azimuth gate of Cao et al. [13] was reimplemented in the same way and is run with its own acceptance test, the association ratio on a subsequent frame, which is not a track-level accuracy. Thresholds and conventions that the papers do not specify were fixed before the runs and are recorded with the code. Equivalence between these reimplementations and the authors' implementations has not been established, so those runs describe the reimplementations on this site and are not used to rank the methods by accuracy. The quantitative evaluation and ablations in this paper assess the proposed method. Section 7.9 distinguishes the requirements of full calibration methods from those of the correspondence stage studied here.

7. Results

7.1. Accuracy on the Labelled Recordings

Table 2 gives the correspondence accuracy on the eight labelled recordings. The consensus assigns 656 of the 672 radar tracks with a counterpart to the correct vehicle, 0.956 to 1.000 per recording and 0.976 in total, with a pair-level precision of 0.951 to 1.000 and a recall of 0.766 to 0.944, from 64 to 123 radar–camera tracklet pairs per recording. A labelled pair scores above $S = 0.5$ for only 0.527 to 0.649 of the tracks (cues only in Table 2), so a threshold on the cues alone would leave more than a third of them unresolved; the consensus decides them by geometry, since the cues enter the assignment cost only as a secondary term. No manual initialization, calibration, or label enters the proposed procedure; the labels are used only to score it.

One further recording tests another day. It was made on 4 August 2023, two months before R0 to R7, with the camera aimed about 13 to 15 degrees lower, and it is the longest recording of the unit from another day with uninterrupted streams (345 s, 44 radar and 150 camera tracks). It is kept apart from the pooled rows as R8 and was labelled with the blind checklist of Section 6 by the first author and by a second person, without seeing the output; the two label sets differ on three radar tracks. With every parameter unchanged the method assigns 41 radar tracks to the correct vehicle against either label set, an accuracy of 0.953 and 0.932, with a pair precision of 0.976. The selected hypothesis is two tracks short of the most accurate of its 67 before refinement, and the cost per radar track is 8.4, above the 3.0 to 4.7 of the eight recordings (Section 7.2), because 107 of the 150 camera tracks have no radar partner and each is paid for at c_0; a usability threshold on the cost would therefore have to be normalized by the tracks of both sensors.

Table 2. Correspondence accuracy on the eight labelled recordings. Accuracy is correct-track coverage among radar tracks with a counterpart, counting unmatched tracks as unsuccessful; precision and recall are per radar–camera tracklet pair; pairs: tracklet pairs assigned; cues only: fraction of radar tracks whose labelled counterpart scores above $S = 0.5$. The pooled rows are computed from the per-recording counts; R0 was used during development. Bold marks the pooled results.

			Proposed Consensus			
Rec.	**Labelled**	**Cues Only**	**Accuracy**	**Precision**	**Recall**	**Pairs**
R0	74	0.649	0.973	0.977	0.944	86
R1	100	0.580	0.990	1.000	0.836	112
R2	114	0.596	0.956	0.951	0.880	123
R3	80	0.550	0.963	1.000	0.925	86
R4	96	0.583	0.979	0.991	0.766	106
R5	55	0.527	1.000	1.000	0.928	64
R6	82	0.646	0.988	0.988	0.857	85
R7	71	0.606	0.972	0.987	0.926	76
R1–R7	598	0.587	**0.977**	**0.986**	**0.863**	**652**
All	672	0.594	**0.976**	**0.985**	**0.872**	**738**

7.2. Self-Scoring Property of the Assignment Cost

The selection criterion of Section 5 is the total assignment cost, which uses no reference data. Figure 5 plots it against the accuracy of the matching that each seed hypothesis produces, for all 71 to 89 hypotheses of every recording, and Table 3 summarizes the consensus stage. The cost correlates with the accuracy at $r = -0.71$ to -0.94 on the eight recordings (Spearman rank correlation -0.71 to -0.92). Most hypotheses are wrong: the median accuracy over the hypotheses is 0.16 to 0.31, because a hypothesis is correct only if both of its seeds are true pairs and only four to six of the fifteen seeds are. The wrong hypotheses are wrong in different ways and their costs scatter over a wide range, so the relation is not a straight line and the correlation coefficient understates how reliably the minimum is placed. The lowest-cost hypothesis is the most accurate of all hypotheses on seven recordings, with an accuracy of 0.93 to 1.00 before any refinement, and on R2 it is the second most accurate (0.912 against 0.921 for the hypothesis that costs 461 instead of 460). On every recording the few lowest-cost hypotheses form a cluster at the top left of the plot; they share true seeds and differ only in which of the marginal tracks they assign.

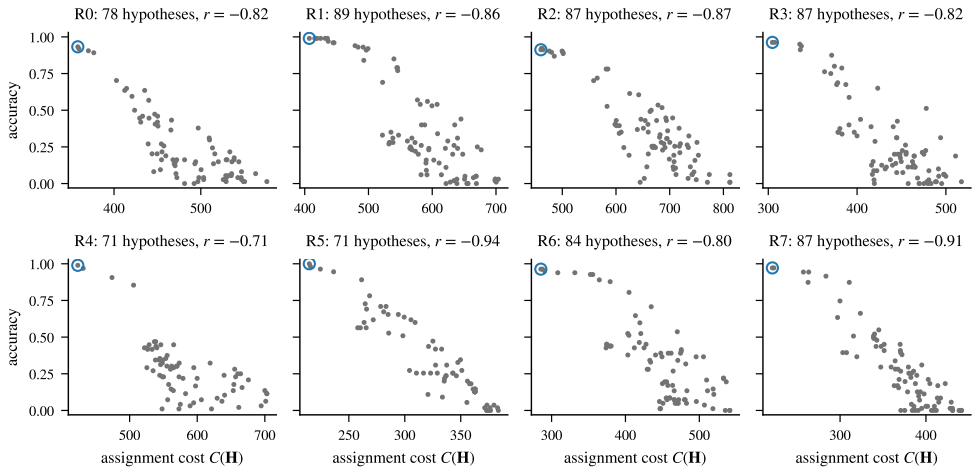


Figure 5. Assignment cost $C(\mathbf{H})$ against the accuracy of the matching that each seed hypothesis produces, for all admissible hypotheses of the eight recordings; the selected (lowest-cost) hypothesis is circled. The accuracies are known only because the recordings are labelled; the cost is computed from the data alone.

Table 3. The consensus stage on the eight recordings. Upper block, hypothesis generation and selection. Overlapping: radar–camera pairs with common frames; vetoed: pairs removed by the direction veto; true seeds: true pairs among the 15 seeds; hyp.: number of hypotheses; r: correlation of cost and accuracy over the hypotheses; lowest-cost: accuracy of the selected hypothesis before refinement, and the rank by cost of the most accurate hypothesis; median: median accuracy over all hypotheses. Lower block, cost and refinement. C: cost of the selected and of the second-best hypothesis and the selected cost per radar track; refined: accuracy after the two refinement rounds; attached: camera tracklets added by the attachment step; d: median distance under the refined homography between the projected radar track and the camera track, for labelled pairs and for all other overlapping pairs.

						Lowest-Cost		
Rec.	**Overlapping**	**Vetoed**	**True Seeds**	**Hyp.**	r	**Acc.**	**Rank**	**Median**
R0	916	220	4/15	78	−0.82	0.932	1	0.16
R1	1015	209	6/15	89	−0.86	0.990	1	0.26
R2	1316	317	6/15	87	−0.87	0.912	2	0.31
R3	1002	230	5/15	87	−0.82	0.963	1	0.19
R4	1189	300	4/15	71	−0.71	0.990	1	0.25
R5	557	182	4/15	71	−0.94	1.000	1	0.26
R6	722	198	6/15	84	−0.80	0.963	1	0.30
R7	753	229	5/15	87	−0.91	0.972	1	0.21

	C(H)					
Rec.	**Best**	**2nd**	**Per Track**	**Refined**	**Attached**	d **True/Other [px]**
R0	358	360	4.65	0.973	12	7.0/197
R1	407	416	3.92	0.990	13	9.0/213
R2	460	461	3.65	0.939	10	8.6/206
R3	305	307	3.71	0.963	9	9.3/213
R4	423	431	4.23	0.979	11	8.3/207
R5	215	216	3.77	1.000	9	7.4/209
R6	285	288	3.21	0.988	3	9.0/221
R7	221	223	3.03	0.972	6	8.4/202

Refinement on all matched pairs then changes the accuracy of the selected hypothesis by −1.0 to +4.1 points (+4.1 on R0, +2.6 on R2, +2.4 on R6, −1.0 on R4), and the attachment step adds 3 to 13 tracklets per recording, which raises the recall and, where a second tracklet is the one that reaches the labelled counterpart, the accuracy (R2: 0.939 to 0.956).

The absolute level of the cost at the selected hypothesis, divided by the number of radar tracks, is 3.0 to 4.7 on all eight recordings, which is the level at which most tracks are matched within a few pixels and paid for mainly by their similarity term. A recording with too little shared traffic or degraded tracks would be expected not to reach it, so the level could serve as a usability check, once normalized by the tracks of both sensors (Section 7.1), but the present data contain no such recording and the check remains unvalidated. The geometric separation the selected model produces is large: under the refined homography, the labelled pairs project within a median of 7 to 9 pixels of their camera track, against 197 to 221 pixels for all other overlapping pairs (last column of Table 3).

7.3. Parameter Sensitivity of the Consensus Stage

The consensus stage has nine parameters, listed in Section 5. They were fixed on recording R0 during the development of the method, before the labels of R1 to R7 existed, and were not changed afterwards. They were set by hand from the quantities they gate rather than by a search: σ at the scale of the error with which a two-seed homography places the far end of a track, tens of pixels; c_0 so that, by Equations (8) and (9), a pair is accepted when its projected distance is about σ and its similarity above about 0.2, or its distance about 2σ and its similarity above about 0.7; the seed spread at about a lane

width so that two seeds from the same lane are not combined; thirty common frames, two to four seconds, so that a seed pair determines the homography with margin over the four-point minimum; and fifteen seeds so that, at the precision the cues reach, several correct hypotheses are generated (Table 3). Each of them was varied in turn over a wide range, with the others at their values, and accuracy, precision, and recall were recorded on all eight recordings. Figure 6 shows the five parameters that change the result; the other four, the seed score threshold (0.4 to 0.6), the minimum number of common frames per seed (15 or 30), the minimum seed spread (1.5 to 6 m), and the number of refinement rounds (0 to 4), leave accuracy, precision, and recall within one point of their values or are described below.

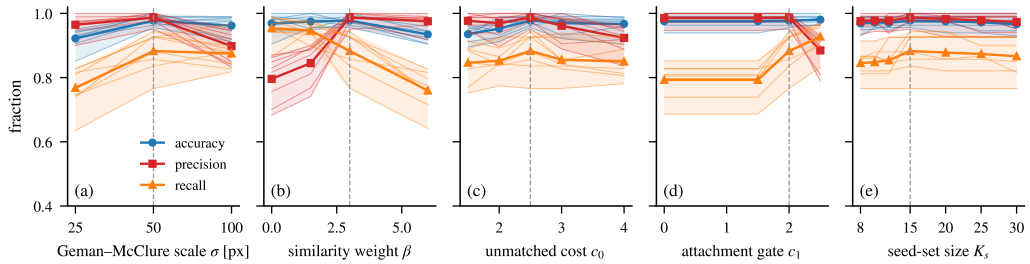


Figure 6. Sensitivity of the correspondence to the parameters of the consensus stage that change the result, one at a time, with the others at their values (dashed line): (**a**) Geman–McClure scale σ; (**b**) similarity weight β; (**c**) unmatched cost c_0; (**d**) attachment gate c_1; (**e**) seed-set size K_s. Thick lines: mean over the eight recordings; thin lines: the individual recordings; shaded bands: the range between the lowest and the highest recording. Accuracy is per radar track; precision and recall are per tracklet pair.

The size of the seed set matters in one direction. With eight seeds the lowest-cost hypothesis is worse on average (0.85 against 0.97 with fifteen) and reaches 0.40 on R2, where only three of the eight seeds are true pairs, but refinement and attachment recover most of the loss and the final accuracy stays at 0.97 on average and 0.93 at worst. With twenty to thirty seeds the number of hypotheses grows to 144 to 336 and the mean accuracy falls slowly to 0.97, because more low-cost wrong hypotheses compete. Twelve to twenty seeds are equivalent.

The parameters of the assignment cost trade precision against recall. The Geman–McClure scale σ decides how far a projected track may lie from a camera track before it is treated as a gross mismatch: at 25 px the recall falls to 0.77 and the accuracy to 0.92, because the two-seed homography places the far end of many tracks more than 25 px off and those tracks are left unmatched; at 100 px the precision falls to 0.90, because neighbouring lanes become matchable. The similarity weight β behaves in the same way. With $\beta = 0$, i.e., geometry alone, the recall rises to 0.95 but the precision falls to 0.80, since without the time and speed cues a track is matched to whichever camera track its projection happens to touch; with $\beta = 6$, the cues dominate the geometry and the recall falls to 0.76. The unmatched cost c_0 sets where a track is left unassigned: below 2.5 too many tracks are left out (accuracy 0.94 at $c_0 = 1.5$), above it wrong tracks are forced (precision 0.92 at $c_0 = 4$). The accuracy per radar track stays above 0.92 over the whole range of every one of these parameters; what moves is how many marginal tracklets are attached and how many wrong ones are admitted.

The geometry weight α behaves as the mirror image of β (Figure 7, left): at $\alpha = 1$ the cues dominate and the precision falls to 0.77 while the recall stays at 0.85, at $\alpha = 10$ the geometry dominates and the recall falls to 0.78 at a precision of 0.96, and the accuracy is 0.92 at both ends and highest at the value used. Only the ratios matter: scaling α, β, c_0, and c_1 together by two reproduces every number of Table 2 exactly, since the cost is homogeneous

in them. The nominal vehicle width of the speed cue (Figure 7, right) changes the accuracy by at most 1.2 points and the precision by at most 3 points between 1.5 and 6.5 m, although the fraction of labelled pairs that the cues alone place above $S = 0.5$ rises from 0.58 to 0.69 over that range, consistent with the factor of about 0.3 measured in Section 4; the decision is made by the geometry. Replacing the four-state quantization of the speed cue by eight states or by the continuous score $1 - \min(1, |v_r - v_c|/1.5)$ lowers the mean accuracy by about one point (0.968 and 0.969 against 0.978) and the recall by three to five points, so the quantization is kept.

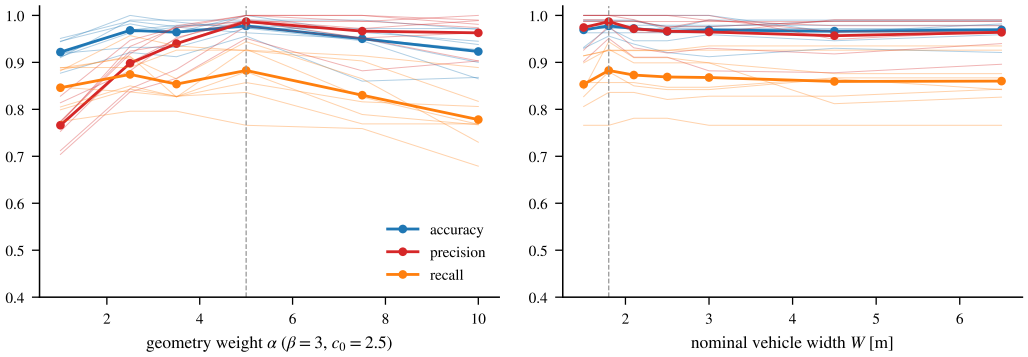


Figure 7. Sensitivity to the geometry weight α (left; $\beta = 3$, $c_0 = 2.5$) and to the nominal vehicle width W of the speed cue (right), in the format of Figure 6; the dashed line marks the value used.

The refinement and attachment steps behave as designed. Without refinement the accuracy is 0.97 and the recall 0.86 (against 0.98 and 0.88 with two rounds); a fourth round lowers the accuracy to 0.96 on average and to 0.89 on R0, because the refit on a larger and slightly contaminated pair set starts to admit marginal pairs. Without the attachment step, the accuracy is unchanged but the recall falls to 0.79, which is the fragmentation of Section 7.8; with a looser gate ($c_1 = 2.5$), the recall rises to 0.93 at a precision of 0.89. The values used, $c_1 = 2.0$ and two rounds, are the point where the precision is preserved.

7.4. Reference Height and the Plane of the Homography

Section 3 argued that the fitted homography is that of the plane of the vehicle reference points and that only the variation of their height from vehicle to vehicle survives the fit. Two measurements examine this. First, the signed residual between the projected radar position and the box centre under the refined homography, averaged over the common frames of each labelled pair, has a median of -4.5 to $+2.2$ pixels horizontally and -2.2 to $+1.8$ pixels vertically on the eight recordings, with interquartile ranges from 6 to 13 pixels, as it must, since the homography is fitted to these pairs. The residual grows with the size of the box: on R2, the per-pair mean distance has a median of 4.7 pixels for the smallest third of the boxes (26 pixels high, median range 56 m), 10.0 pixels for the middle third and 13.0 pixels for the largest third (63 pixels high, 32 m), and on R0 it is 3.5, 7.1, and 12.7 pixels; the vertical component is 0.6, 3.0, and 4.5 pixels on R2 and -0.3, -1.9, and -6.2 pixels on R0. Large boxes are both nearer and taller, so this is the height effect of Section 3 together with the pixel scale of the near field, and its size is the 5 to 13 pixels that the geometric estimate predicts. Second, a compensation was applied: the bottom centre of the bounding box, which lies close to the road contact point, replaced the box centre as the camera reference in the cues and in the geometry. The accuracy falls from 0.978 to 0.962 on average (0.904 to 0.990 per recording; R2 from 0.956 to 0.904), the precision from 0.987 to 0.962, and the recall from 0.883 to 0.832, and the median distance of the labelled pairs under the refined homography moves from 7.0 to 9.3 pixels to 7.0 to 10.4 pixels. Moving

the camera reference to the road plane therefore removes no systematic shift, because none common to all vehicles exists, and it costs accuracy. The box centre is kept.

7.5. Time Offset Between the Streams

The camera time base was shifted against the radar by whole frames, -5 to $+5$ frames or up to 0.37 to 0.65 s at the frame intervals of the three presets, and the whole stage was rerun on the eight recordings (Figure 8). The mean accuracy stays within 0.5 points of its value for shifts of $+1$ and $+2$ frames (0.975 and 0.973 against 0.978) and falls to 0.959 at -1, 0.962 at $+3$, 0.91 to 0.92 at -2 and -3, 0.90 at $+5$, and 0.80 at -5; the precision follows the same curve, and the recall falls faster, from 0.883 to 0.82 at -1 and to 0.67 to 0.76 at ± 5. The response is asymmetric: shifting the camera later than the radar (positive shifts, which pair a radar frame with an earlier image) costs less than shifting it earlier, consistent with a residual offset of that sign, since the camera is triggered by the radar frame and its image lags the radar measurement by the radar's processing latency (Section 4). The curve is centred near a shift of $+1$ frame, in line with the offset of one to two frame intervals that the calibration stage estimates (Section 4): the accuracy changes by at most half a point between shifts of 0 and $+2$, falls by about two points one frame further on either side, and degrades beyond that, when the time overlap and the speed states of a true pair no longer line up and the two-seed homographies are fitted to displaced samples. A method that needs synchronized streams would have to obtain this alignment from elsewhere; here, the frame-level pairing of Section 4 suffices.

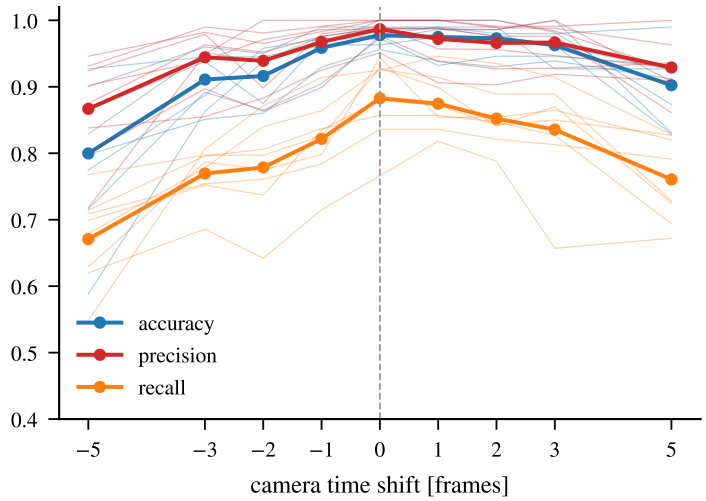


Figure 8. Sensitivity of the correspondence to a shift of the camera time base by whole frames (one frame is 74 to 129 ms depending on the preset), in the format of Figure 6.

7.6. Cost of a Lane-Shifted Hypothesis

To test the lane-shift failure mode of Section 8 directly, the refined homography of each recording was composed with a translation of the radar plane across the dominant lane direction by one and by two lane widths, 3.8 and 7.6 m, the spacing of adjacent lane centres measured from aerial imagery of the site, and the cost and the assignment were recomputed under the shifted model with everything else unchanged. In whichever direction costs less, one lane over, the total cost rises from 213 to 462 under the refined model to 335 to 694, or per radar track from 3.0 to 4.3 to 5.1 to 6.2, and two lanes over to 6.0 to 7.1 per track; the margin between the selected hypothesis and its cheaper one-lane copy is 120 to 230 cost units on every recording, against 1 to 9 between the selected and the second-best of the generated hypotheses (Table 3). Under the one-lane shift, the labelled pairs project a

median of 44 to 73 pixels from their camera tracks instead of 7 to 9, and 49 to 83% of the radar tracks are still assigned to the correct vehicle, depending on the recording and the direction, because a shift of about σ costs only part of the gate; but fewer pairs are matched on every recording and each unmatched track is paid for at c_0. Two lanes over, the accuracy is at most 0.36 and 24 to 57 pairs are matched among 57 to 126 radar tracks. A shifted hypothesis is thus rejected by the cost rather than by the gate, and by a margin far wider than the separation among the generated hypotheses, on every recording of this site; what the cost cannot promise is the same margin on a road whose lanes carry the same traffic at the same times.

7.7. Cue Ablation

One cue at a time was removed, the speed profile, the direction veto, the graded direction score, and the turning-event score, on all eight recordings. The time overlap is not removed, because every cue is computed on the frames common to the two tracks, so it defines the candidate set rather than ranking it. Over the eight recordings 656 radar tracks are correct with all cues, 651 without the speed profile, 652 without the turning-event score, 651 without the veto and 650 without the graded direction score. Without the speed profile R2 loses 3.5 points of accuracy (0.956 to 0.921, with eleven wrong pairs instead of six) and the recall falls on all eight recordings, from 0.883 to 0.828 on average, so the cue contributes as much as the others although its state agreement separates labelled pairs from the other overlapping pairs only weakly (Section 4). Without the turning-event score, the recall falls on six of the eight recordings (by 0.01 to 0.15, most on R5, whose curved paths through the intersection are where the cue separates candidates), is unchanged on R3, and rises by 0.007 on R4 at the price of one more wrong pair. Without the graded direction score the recall falls on seven of the eight and the accuracy on R0, R2, and R5 (to 0.959, 0.921, and 0.982). Without the veto R0, R1, R3, R4, and R7 are unchanged, R2 loses 3.5 points of accuracy (0.956 to 0.921), and R5 and R6 one to two points of accuracy or precision, because the Doppler gate of the radar tracker has removed most of the direction conflicts the veto used to catch, but not all of them. Each cue therefore contributes a few pairs on most recordings and a few points of accuracy on the recordings where traffic is densest; the decision of which vehicle a radar track belongs to is made by the geometry.

7.8. Failure Analysis

Table 4 sorts every disagreement with the labels into its cause, and Figure 9 shows the final matching of recording R2, the recording with the most wrong tracks. Of the 672 radar tracks with a counterpart, 656 are correct, 5 are assigned to a wrong camera track, and 11 receive no camera track; of the 36 radar tracks without a counterpart, 4 receive one. The five wrong tracks are on R2 (3), R4, and R7. In each case the camera tracklet belongs to the vehicle immediately ahead or behind in the same lane, which has the same timing, speed, and direction and projects a few metres along the lane from the radar track so that under the gate c_0 the wrong tracklet costs less than leaving the track unmatched. The eleven missed tracks are mostly short radar tracks whose only camera counterpart is a long tracklet that overlaps them for a small fraction of its duration, so the time-overlap score and hence the similarity term is small and the pair is left unassigned. The four spurious assignments are radar tracks the labeller judged to have no valid counterpart.

At the pair level one cause dominates. Of the 834 labelled pairs, 727 are assigned and 107 are missed; 90 of the 107 are the second tracklet of a radar track whose first tracklet was assigned correctly, 6 belong to the five wrongly assigned tracks, and 11 to the missed tracks. The recall of 0.77 on R4, the lowest in Table 2, is entirely of this kind: R4 has 41 radar tracks with two labelled tracklets, more than any other recording, and 30 second tracklets are not

attached, while 94 of its 96 tracks are correct. The attachment step recovers 71 of the 162 second tracklets over the eight recordings. The ones it misses are short tracklets at the far end of the approach, where the oblique view compresses the lane into a few pixel rows and the refined homography, fitted to whole tracks, projects the radar track a few pixels off, or tracklets that begin after an occlusion and overlap the radar track for a few frames only. The precision is unaffected: only two wrong tracklets are attached to correctly assigned tracks over all eight recordings. For the downstream estimation, which needs correct pairs rather than all pairs, this is the right side to err on.

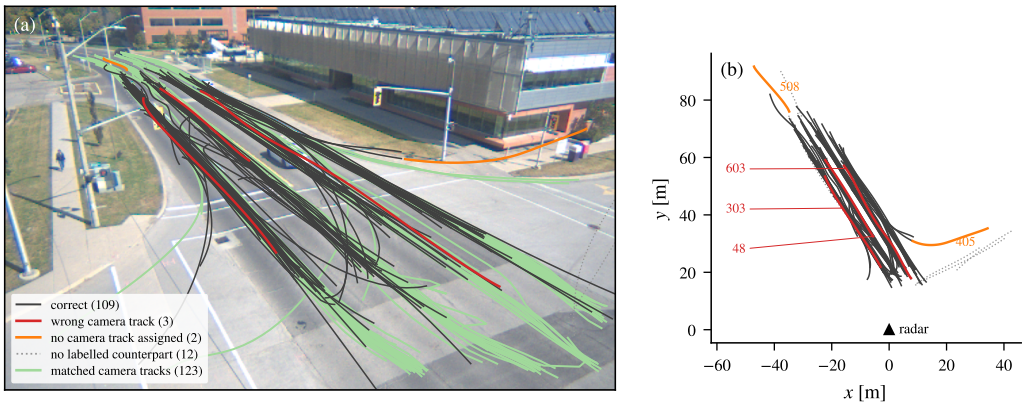


Figure 9. The final matching on recording R2 (126 radar tracks, 204 camera tracks). (**a**) Every radar track projected into the image through the refined homography, coloured by its outcome against the labels: correct (109, dark grey), assigned to a wrong camera track (3, red), left without a camera track (2, orange), and radar tracks without a labelled counterpart (12, dotted); the 123 matched camera tracks are drawn underneath in light green. (**b**) The same tracks in the radar frame, with the identifiers of the wrong and missed tracks. The three wrong tracks received the camera track of the vehicle immediately ahead or behind in the same lane.

Table 4. Every disagreement with the labels by cause. Tracks: radar tracks with a counterpart that are correct, assigned to a wrong camera track, or left without one; "none" tracks are those labelled without a counterpart and "spurious" those among them that received a camera track. Pairs: labelled tracklet pairs not assigned, split into second tracklets of a correctly assigned radar track, pairs of a wrongly assigned track, and pairs of a missed track; extra: additional wrong tracklets attached to correctly assigned tracks.

	Radar Tracks with a Counterpart				None		Labelled Pairs				
Rec.	**Total**	**Correct**	**Wrong**	**Missed**	**Tracks**	**Spurious**	**Total**	**Second Tracklet**	**Of Wrong**	**Of Missed**	**Extra**
R0	74	72	0	2	3	2	89	3	0	2	0
R1	100	99	0	1	4	0	134	21	0	1	0
R2	114	109	3	2	12	1	133	11	3	2	2
R3	80	77	0	3	2	0	93	4	0	3	0
R4	96	94	1	1	4	0	137	30	1	1	0
R5	55	55	0	0	2	0	69	5	0	0	0
R6	82	81	0	1	7	1	98	13	0	1	0
R7	71	69	1	1	2	0	81	3	2	1	0
All	672	656	5	11	36	4	834	90	6	11	2

7.9. Published Methods: Information Requirements and a Limited Comparison

Table 5 compares the information required by the published calibration methods with the requirements of the correspondence stage developed here. Since the outputs differ, the table describes the information needed at each stage and does not rank calibration accuracy.

Table 5. Information requirements and modelling assumptions of the published methods. "Not supplied" refers to an external input, not to a numerical initialization computed from the observations. Entries distinguish supplied information from quantities recovered from data. The proposed column describes only the correspondence stage of this paper; the other columns describe complete calibration methods. This is a comparison of requirements, not an accuracy ranking.

Information/Assumption	Liu et al. [12]	Cao et al. [13]	Zhu et al. [14]	Proposed
Initial radar–camera rotation/translation	Not supplied	Not supplied	Not supplied	Not supplied
Camera intrinsics	Calibrated intrinsics supplied	Calibrated intrinsics supplied	Intrinsics supplied	Not required for correspondence
Installation height	No separate height input	Height enters BEV conversion	Road plane estimated from radar	Not required for correspondence
Lane information	No lane map required	Heading obtained from traffic	Lane geometry extracted; marking dimensions and width priors	No lane map or marking lengths required
Mounting information	No relative-pose prior	Approximate horizontal co-alignment assumed	No relative-pose prior	No relative-pose prior
Temporal information	Aligned timestamps; interpolation	Corresponding sensor frames	Synchronized frames	Common time base; overlapping tracks
Road representation	Ground-plane representation	Ground plane in BEV	Radar plane estimated by RANSAC	Common plane with non-collinear support
Traffic evidence	Paired moving trajectories	Spatially separated pairs; shared target visibility	Lane-change and road-entry/exit events	Several sufficiently extended vehicle trajectories
Output	Extrinsic calibration	Extrinsic calibration	Extrinsic calibration	Track correspondences and a plane-to-image mapping

None of the three methods requires a supplied initial radar–camera extrinsic matrix. Liu et al. use known camera intrinsics and interpolate timestamped trajectories. Cao et al. additionally exploit approximate horizontal alignment of the integrated sensors; their BEV conversion uses installation height and a lane heading obtained from traffic. The height entry records the quantity appearing in their Equation (1), rather than implying that the paper prescribes a separate manual survey. Zhu et al. extract lane geometry from images and use known physical marking lengths and spacing to recover longitudinal distance; this information may come from road-marking standards rather than a site survey. Their event detection also uses road- and vehicle-width assumptions, and event association uses synchronized observations. Thus, automatic extraction of lane lines or a road plane should be distinguished from supplying metric scene information.

The proposed correspondence stage requires temporally overlapping trajectories on a common surface with enough spatial support to determine a homography. It does not require supplied camera intrinsics, an initial extrinsic transform, lane markings, or a metric installation height. It nevertheless needs a common time base to evaluate motion cues and overlap; independence from a particular hardware synchronization mechanism is not independence from temporal alignment. Its output is a set of track correspondences and a plane-to-image mapping. Recovering metric extrinsics is a separate stage, which needs a camera model and the geometry of the mount that the correspondence does not. The component ablation in Section 7.7 assesses the contributions within the proposed correspondence procedure.

Table 6 gives the result of the reimplemented association steps of Section 6 on the same tracks and labels, with the proposed consensus beside them. The per-pair PnP of Liu et al. assigns 137 of the 672 radar tracks to the correct vehicle in its published representation and 132 in the adapted one, from 320 and 315 returned pairs, so that two of every five returned pairs are correct; the event matching of Zhu et al. returns 39 and 73 pairs, 28 and 36 of them correct, a coverage of 4 to 5% at a precision of 0.72 and 0.49. Coverage and precision have different denominators and must be read together: a method that returns

few pairs can have a low coverage and a high precision at once, as the event method does on R0, where five of its six pairs are correct. The azimuth gate of Cao et al. passes its own acceptance test, an association ratio above 0.9 on a subsequent frame, on R2 and R5 in the published representation and on no recording in the adapted one; where it passes, the 25-frame confirmation that follows identifies 12 pairs on R2, 11 of them correct, and 18 on R5, all correct and covering 16 distinct radar tracks, and no run reaches a validated fine update. These numbers describe documented reimplementations on one site. Choices the papers leave open, the validation track of Liu et al., the thresholds of the azimuth gate, the lane clustering and the event segments of Zhu et al., were fixed before the runs and are material to the result, and agreement with the authors' implementations has not been established, so the table does not rank the methods; it records what each association step, as described, returns on lane-keeping traffic seen from this unit, which is the condition that Section 3 argues a per-pair decision cannot resolve.

Table 6. Association steps of the reimplemented methods on the eight labelled recordings, against the same labels as Table 2. Cells give the radar tracks assigned to the correct vehicle, with the returned pairs and the pair precision in parentheses; "published" and "adapted" are the two representations of Section 6. Cao et al.: whether the conditional coarse acceptance test of the method passes in the published representation (it passes on no recording in the adapted one) and the pairs its 25-frame confirmation identifies where it does. The reimplementations follow the papers under choices they leave open; the table describes them on this site and does not rank the methods. The column of the proposed consensus is set in bold.

		Liu et al. [12]		Zhu et al. [14]		Cao et al. [13]	Proposed
Rec.	**Labelled**	**Published**	**Adapted**	**Published**	**Adapted**	**Coarse Gate**	**Consensus**
R0	74	19 (44, 0.43)	12 (36, 0.33)	5 (6, 0.83)	4 (7, 0.57)	fails	**72 (86, 0.98)**
R1	100	20 (53, 0.38)	15 (55, 0.27)	1 (2, 0.50)	1 (3, 0.33)	fails	**99 (112, 1.00)**
R2	114	43 (70, 0.61)	31 (71, 0.44)	8 (11, 0.73)	8 (14, 0.57)	passes, 11/12 pairs	**109 (123, 0.95)**
R3	80	8 (47, 0.17)	9 (44, 0.20)	4 (7, 0.57)	4 (14, 0.29)	fails	**77 (86, 1.00)**
R4	96	23 (42, 0.55)	22 (41, 0.54)	5 (6, 0.83)	11 (19, 0.58)	fails	**94 (106, 0.99)**
R5	55	1 (18, 0.06)	5 (6, 0.83)	0 (0, –)	0 (2, 0.00)	passes, 18/18 pairs	**55 (64, 1.00)**
R6	82	13 (23, 0.57)	22 (34, 0.65)	3 (4, 0.75)	5 (8, 0.63)	fails	**81 (85, 0.99)**
R7	71	10 (23, 0.43)	16 (28, 0.57)	2 (3, 0.67)	3 (6, 0.50)	fails	**69 (76, 0.99)**
All	672	137 (320, 0.43)	132 (315, 0.42)	28 (39, 0.72)	36 (73, 0.49)	passes on 2 of 8	**656 (738, 0.99)**

8. Discussion and Limitations

The claims of this paper are bounded by six limitations, taken in turn below: the single site and day, the plane assumption, the recall bound set by the fragmentation of camera tracks, the relative nature of the assignment cost, the labelling protocol, and the comparison with published methods, which rests on reimplementations.

All eight recordings come from one unit above one intersection approach, made on one day. They differ in preset, in traffic, and in the tracks they contain, and the 672 labelled radar tracks are, to the authors' knowledge, the largest correspondence reference reported for a roadside radar–camera pair, but the road geometry, the mounting height, and the lens are the same throughout. We investigated the public roadside data set V2X-Radar-I [34] for external evaluation. In the sequences examined, the lower sensor mounting height limited the visible extent of vehicle trajectories, leaving insufficient trajectory support for reliable extraction of the motion cues. These sequences were therefore not included in the quantitative evaluation. The method requires sufficiently extended trajectories for its motion cues and non-collinear spatial support for its homography hypotheses; its performance across mounting heights and sites remains untested. The recording from

another day and camera aim of Section 7.1 is the one piece of evidence beyond that day: with every parameter unchanged the method assigns 41 radar tracks to the correct vehicle against either of two label sets, an accuracy of 0.93 to 0.95. What would change elsewhere can be stated from the method. A lower mounting or a longer focal length shortens the visible extent of the trajectories, which is the condition that failed on V2X-Radar-I, and an approach without turning traffic or a second lane removes the lateral structure the seed spread requires. Of the parameters, those in pixels, the inlier threshold and σ, scale with the image resolution and with the projection error of a two-seed homography, which grows with the depth range of the scene; the seed spread is a lane width; the speed thresholds are in metres per frame interval and follow the frame rate; the dimensionless weights α, β, c_0, and c_1 and the seed count have no site dependence that we know of, and the sweep of Section 7.3 keeps the accuracy above 0.92 over the ranges swept there.

The consensus relies on a plane. The homography model assumes that the vehicle reference points lie on one plane visible to both sensors, which holds on the approach studied here and on most urban and highway segments, and which the cost of the selected hypothesis confirms on each recording. On strongly crowned, banked, or multi-level roads, or where an elevated ramp crosses the field of view, no single homography would explain all tracks. The method would be expected to report a high minimum cost and a poor assignment rather than a confident wrong one, which has not been tested, and it would not associate such a scene. Vehicle height above the plane is not a problem for the correspondence, because it moves all projected points of a track by a similar amount and the Geman–McClure cost tolerates it, but it is one reason the homography is a coarse model and not the calibration.

Recall is bounded by the fragmentation of the camera tracks. Of the 107 labelled pairs the method misses, 90 are the second tracklet of a radar track whose first tracklet was assigned correctly. The attachment step recovers a little under half of the second tracklets (71 of 162); the rest fail its gate because a short tracklet at the far end of the approach, or one that starts after an occlusion, projects too far from the radar track under the two-seed homography refined on whole tracks. A segment-level association, which would match parts of tracks rather than whole tracks, would recover most of these pairs and would also handle the camera tracklets that jump to the vehicle ahead, the one failure the correspondence passes on to the estimation stage. The precision figures, which are what the estimation stage depends on, are not affected.

The assignment cost is a relative criterion within a recording. It ranks hypotheses, and its correlation with accuracy is strong on every recording. Its absolute value per track is also similar across recordings (3.0 to 4.7 at the selected hypothesis), which makes a threshold on it, normalized by the tracks of both sensors (Section 7.1), a plausible usability test, but the present data contain no recording on which the correspondence failed, so the threshold has not been exercised. The cost also cannot detect a matching that is globally consistent but wrong by a lane shift for every track, which would require a wrong homography that fits all tracks at once. Geometrically, a lane shift is a translation of the plane by one lane width across the lanes, and a translation composed with the true homography is again a homography, so the model class contains it. It attains a low cost only if every radar track finds, one lane over, a camera track that coincides with it in time, speed, and direction, since the similarity term and the gate are paid per track; on the approach studied here, the lanes carry different traffic, the outer lane maps onto the kerb and the turning paths through the intersection map onto no lane at all. Section 7.6 constructs the shifted hypothesis and finds its cost 120 to 230 units above the selected one on every recording, against a margin of 1 to 9 between the two best generated hypotheses. A site with one straight lane per direction and dense, uniform traffic would offer less protection. A lane-width prior would

not restore it, because a shift by one lane leaves the lane spacing unchanged; the cost would need an absolute check, for instance, the decomposition of the selected homography with the nominal focal length, which returns the position of the radar relative to the camera and would place a shifted solution a lane width away from a radar that is mounted within a fraction of a metre of the camera. That check has not been implemented.

The automatic assignment was visible during the initial labelling, which introduces a risk of confirmation bias, and re-examining disagreements under the protocol of Section 6 does not eliminate it. The blind relabelling of the 126 radar tracks of R2 by an independent person (Section 6) measures it on that recording: the two label sets agree on 123 of the tracks, and the accuracy of the method on R2 is 0.947 against the independent labels and 0.956 against the paper's. The difference is one radar track; the other seven recordings were not relabelled. The failures that remain after re-examination (Table 4) have identifiable causes. Labelled recall also depends on which short tracklets a human could confirm, so the recall figures are conservative in one direction and the pair counts in the other.

The literature comparison establishes differences in required inputs and modelling assumptions, and the limited comparison of Section 7.9 measures what the association steps of two published methods, reimplemented from their papers, return on the same recordings. Neither demonstrates numerical superiority over the published methods: equivalence to the authors' implementations has not been established and several choices the papers leave open are material, so the measured coverages describe the reimplementations and not the methods as their authors run them. The quantitative evidence here concerns the proposed method's labelled correspondence results, component ablations, and parameter sensitivity on one site. Under identical assumptions, calibrated intrinsics and synchronized streams, the methods differ in what one track is asked to determine. The per-pair PnP of Liu et al. fits a six-degree-of-freedom pose to a single nearly straight track, which Section 3 shows to be rank deficient on lane-bound traffic, so its validation error cannot separate a true pair from a wrong one by a margin that survives the noise of the samples; the azimuth gate of Cao et al. presumes the vertical co-alignment of the sensors, which an integrated unit satisfies, but it cannot tell apart vehicles that share an azimuth, which are the same-lane neighbours of Section 7.8; the event matching of Zhu et al. needs lane changes, which lane-keeping traffic rarely offers. The consensus asks a track for nothing beyond its position on a common plane and lets the ensemble decide; its price is the plane assumption and the lateral structure the seeds require.

The correspondence is the input of the calibration stage that follows, which estimates the rigid transform, the focal length and the time offset from the matched pairs after a robust rejection of its own. A wrong pair biases that estimate, whereas a missed pair only reduces the amount of data it uses, which is why the precision figures of this paper are the ones that matter downstream.

9. Conclusions

This paper addressed the first decision of any targetless radar–camera calibration from traffic: which radar track and which camera track describe the same vehicle, before a calibration exists. A per-pair pose fitted to nearly linear vehicle trajectories can be weakly constrained, motivating a shared model supported by several tracks. The method presented here scores a candidate road-plane homography by the total cost of a global assignment of all radar tracks to all camera tracks under it, with calibration-free cues to rank the candidate pairs that seed the hypotheses; the lowest-cost hypothesis is refined on its matches and completed by attaching the camera tracklets that occlusion had fragmented.

On eight recordings from an operational unit, labelled by hand against the synchronized video, the method assigns 656 of 672 radar tracks with a counterpart to the correct

vehicle, 96 to 100% per recording, with a pair-level precision of 0.95 to 1.00. The selected hypothesis is the most accurate one on seven recordings and the second most accurate on the eighth, without any reference data. The result is insensitive to the parameters of the consensus stage over wide ranges, to the nominal vehicle width of the speed cue, and to a delay of the camera time base by one to two frames; the similarity cues decide how many tracklets are attached rather than which vehicle a track belongs to; and the 16 tracks that are wrong or missed and the 107 pairs that are missed have identifiable causes, of which the fragmentation of camera tracks is by far the largest. On a recording from another day, with the camera aimed differently, the method with unchanged parameters assigns 93 to 95% of the labelled radar tracks to the correct vehicle against two label sets, with a pair precision of 0.98. An independent blind relabelling of the 126 radar tracks of one recording agrees with the labels used here on 123 of them. The literature comparison identifies the camera, scene, mounting, and timing information required by published calibration methods, and the association steps of two of them, reimplemented from their papers under choices the papers leave open, assign 20 and 4% of the same radar tracks to the correct vehicle in their published representation. The correspondence stage developed here uses a common time base and a common-plane assumption without supplied camera intrinsics, lane markings, or an initial extrinsic transform. These requirements describe its scope; they are not evidence of an accuracy advantage over the published methods, and the reimplementations are not a validated comparison.

Three directions follow from the limitations. A segment-level association would recover the fragmented tracklets that bound the recall, an absolute check of the selected homography against the geometry of the unit would turn the relative cost into a test that can reject a recording (Section 8), and evaluation at a second site would provide evidence beyond the single unit studied here. The structure of the method, a consensus over a shared geometric model that scores itself, is not specific to the radar and camera; it applies wherever two sensors observe the same objects moving on a common surface and no calibration between them is known.

Author Contributions: Conceptualization, methodology, software, validation, investigation, data curation, writing—original draft, Z.R.; supervision, resources, writing—review and editing, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The recordings, the track data and the correspondence labels analysed in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BEV	Bird's-Eye View
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
PnP	Perspective-*n*-Point
RANSAC	Random Sample Consensus
SORT	Simple Online and Realtime Tracking
YOLO	You Only Look Once

References

1. Arnold, E.; Dianati, M.; De Temple, R.; Fallah, S. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1852–1864. https://doi.org/10.1109/TITS.2020.3028424.
2. Wang, T.H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; Urtasun, R. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 605–621. https://doi.org/10.1007/978-3-030-58536-5_36.
3. Zhang, C.; Wei, J.; Dai, J.; Qu, S.; She, X.; Wang, Z. A roadside millimeter-wave radar calibration method based on connected vehicle technology. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 117–131. https://doi.org/10.1109/MITS.2022.3224151.
4. Adl, M.; Guo, X.; Mohammad-Alikhani, A.; Abdi, B.; Ahmed, R.; Emadi, A. Roadside Fisheye Vision for Cooperative Perception in V2I-Assisted Automated Driving. *IEEE Open J. Intell. Transp. Syst.* **2025**, *6*, 1221–1234. https://doi.org/10.1109/OJITS.2025.3603968.
5. Abdi, B.; Rokhi, Z.; Vidal, C.; Emadi, A. Scene-Centric Vehicle Trajectory Prediction at Cooperative Intersection Using Decision-Aware Attention Graph Transformer. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 19322–19333. https://doi.org/10.1109/TITS.2025.3589203.
6. Yao, S.; Guan, R.; Huang, X.; Li, Z.; Sha, X.; Yue, Y.; Lim, E.G.; Seo, H.; Man, K.L.; Zhu, X.; et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Trans. Intell. Veh.* **2024**, *9*, 2094–2128. https://doi.org/10.1109/TIV.2023.3307157.
7. Wang, Y.; Jiang, Z.; Li, Y.; Hwang, J.N.; Xing, G.; Liu, H. RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization. *IEEE J. Sel. Top. Signal Process.* **2021**, *15*, 954–967.
8. Major, B.; Fontijne, D.; Ansari, A.; Teja Sukhavasi, R.; Gowaikar, R.; Hamilton, M.; Lee, S.; Grzechnik, S.; Subramanian, S. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: Piscataway, NJ, USA, 2019; pp. 924–932. https://doi.org/10.1109/ICCVW.2019.00121.
9. Peršić, J.; Petrović, L.; Marković, I.; Petrović, I. Online multi-sensor calibration based on moving object tracking. *Adv. Robot.* **2021**, *35*, 130–140. https://doi.org/10.1080/01691864.2020.1819874.
10. Liang, M.; Yang, B.; Chen, Y.; Hu, R.; Urtasun, R. Multi-task multi-sensor fusion for 3D object detection. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019; pp. 7345–7353.
11. Rokhi, Z.; Emadi, A. Radar-Camera Extrinsic Calibration for Roadside Infrastructure: A Systematic Review. *Vehicles* **2026**, *8*, 137. https://doi.org/10.3390/vehicles8060137.
12. Liu, X.; Deng, Z.; Zhang, G. Targetless Radar–Camera Extrinsic Parameter Calibration Using Track-to-Track Association. *Sensors* **2025**, *25*, 949. https://doi.org/10.3390/s25030949.
13. Cao, L.; Zhu, H.; Gu, T.; Xi, G.; Wang, L.; Luo, J.; Feng, Z. Fusion calib: Azimuth angle and multi frame tracking for online extrinsic radar-camera calibration. *Sci. Rep.* **2025**, *15*, 43231. https://doi.org/10.1038/s41598-025-27309-x.
14. Zhu, B.; Hu, Z.; Lu, Z.; Wen, X. Trajectory-Driven Automatic Extrinsic Calibration for Roadside Radar-Camera Fusion. *IEEE Sens. J.* **2025**, *25*, 15502–15510. https://doi.org/10.1109/JSEN.2025.3548274.
15. Shi, P.; Li, Y.; Ren, H.J.; Xia, R.; Shi, W.K.; Zhang, Y.Y. Automatic radar-camera calibration and fusion for traffic perception. *J. Comput. Sci. Technol.* **2026**, *41*, 415–427. https://doi.org/10.1007/s11390-026-6122-z.
16. Peršić, J.; Marković, I.; Petrović, I. Extrinsic 6DoF calibration of a radar–LiDAR–camera system enhanced by radar cross section estimates evaluation. *Robot. Auton. Syst.* **2019**, *114*, 217–230. https://doi.org/10.1016/j.robot.2018.11.023.
17. Domhof, J.; Kooij, J.F.P.; Gavrila, D.M. An extrinsic calibration tool for radar, camera and lidar. In *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA)*; IEEE: Piscataway, NJ, USA, 2019; pp. 8107–8113. https://doi.org/10.1109/ICRA.2019.8794186.
18. Agrawal, S.; Bhanderi, S.; Doycheva, K.; Elger, G. Static multitarget-based autocalibration of RGB cameras, 3-D radar, and 3-D lidar sensors. *IEEE Sens. J.* **2023**, *23*, 21493–21505. https://doi.org/10.1109/JSEN.2023.3300957.
19. Sengupta, A.; Yoshizawa, A.; Cao, S. Automatic radar-camera dataset generation for sensor-fusion applications. *IEEE Robot. Autom. Lett.* **2022**, *7*, 2875–2882.
20. Wise, E.; Peršić, J.; Grebe, C.; Petrović, I.; Kelly, J. A continuous-time approach for 3D radar-to-camera extrinsic calibration. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: Piscataway, NJ, USA, 2021; pp. 13164–13170.
21. Deng, J.; Hu, Z.; Lu, Z.; Wen, X. FusionCalib: Automatic extrinsic parameters calibration based on road plane reconstruction for roadside integrated radar camera fusion sensors. *Pattern Recognit. Lett.* **2023**, *176*, 7–13. https://doi.org/10.1016/j.patrec.2023.10.015.
22. Houenou, A.; Bonnifait, P.; Cherfaoui, V.; Boissou, J.F. A track-to-track association method for automotive perception systems. In *Proceedings of the 2012 IEEE Intelligent Vehicles Symposium*; IEEE: Piscataway, NJ, USA, 2012; pp. 704–710.

23. Ouaknine, A.; Newson, A.; Rebut, J.; Tupin, F.; Pérez, P. CARRADA dataset: Camera and automotive radar with range-angle-Doppler annotations. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 5068–5075.
24. He, W.; Zhang, G.; Liu, X.; Deng, Z.; Li, X.; Ye, Y. A low-cost temporal-spatial calibration method for radar and camera systems. *Expert Syst. Appl.* **2026**, *306*, 130879. https://doi.org/10.1016/j.eswa.2025.130879.
25. Jin, C.; Zhu, B.; Deng, J.; Hu, Z.; Wen, X. Automatic calibration and association for roadside radar and camera based on fluctuating traffic volume. *Meas. Sci. Technol.* **2024**, *35*, 055118. https://doi.org/10.1088/1361-6501/ad29e6.
26. Du, Y.; Qin, B.; Zhao, C.; Zhu, Y.; Cao, J.; Ji, Y. A novel spatio-temporal synchronization method of roadside asynchronous MMW radar-camera for sensor fusion. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22278–22289. https://doi.org/10.1109/TITS.2021.3119079.
27. Durmaz, O.; Cevikalp, H. Targetless Radar–Camera Calibration via Trajectory Alignment. *Sensors* **2025**, *25*, 7574. https://doi.org/10.3390/s25247574.
28. Huang, G.; Dong, Q.; Cao, H.; Chen, Z.; Hu, Z. Weighted registration of multiple trajectories of dynamic objects for online calibration of MMW radar and camera. *Meas. Sci. Technol.* **2025**, *36*, 065102. https://doi.org/10.1088/1361-6501/add75a.
29. Zhao, C.; Ding, D.; Shi, Y.; Ji, Y.; Du, Y. Graph matching-based spatiotemporal calibration of roadside sensors in cooperative vehicle-infrastructure systems. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 9281–9295. https://doi.org/10.1109/TITS.2024.3374281.
30. Cheng, L.; Cao, S. Online targetless radar-camera extrinsic calibration based on the common features of radar and camera. In *Proceedings of the NAECON 2023—IEEE National Aerospace and Electronics Conference*; IEEE: Piscataway, NJ, USA, 2023; pp. 294–299. https://doi.org/10.1109/NAECON58068.2023.10366051.
31. Cheng, L.; Cao, S. Radar-Camera Fused Multi-Object Tracking: Online Calibration and Common Feature. *IEEE Trans. Intell. Transp. Syst.* **2026**, *27*, 1295–1311. https://doi.org/10.1109/TITS.2025.3624716.
32. Wu, Y.; Fan, J.; Ma, Y.; Huang, L.; Cui, G.; Guo, S. An Automatic Extrinsic Calibration Method for mmWave Radar and Camera in Traffic Environment. *IEEE Trans. Intell. Transp. Syst.* **2026**, *27*, 666–680. https://doi.org/10.1109/TITS.2025.3625727.
33. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 11621–11631.
34. Yang, L.; Zhang, X.; Li, J.; Wang, C.; Ma, J.; Song, Z.; Zhao, T.; Song, Z.; Wang, L.; Zhou, M.; et al. V2X-Radar: A Multi-modal Dataset with 4D Radar for Cooperative Perception. *arXiv* **2025**, arXiv:2411.10962. https://doi.org/10.48550/arXiv.2411.10962.
35. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. https://doi.org/10.1145/358669.358692.
36. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2003.
37. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*; AAAI Press: Washington, DC, USA, 1996; pp. 226–231.
38. Jocher, G.; Qiu, J.; Liu, M.; Lyu, S.; Akyon, F.C.; Kalfaoglu, M.E. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. *arXiv* **2026**, arXiv:2606.03748. https://doi.org/10.48550/arXiv.2606.03748.
39. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2016; pp. 3464–3468.
40. Steinecker, T.; Wuensche, H.J. A Simple and Model-Free Path Filtering Algorithm for Smoothing and Accuracy. In *Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7. https://doi.org/10.1109/IV55152.2023.10186704.
41. Geman, S.; McClure, D.E. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst.* **1987**, *52*, 5–21.
42. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. https://doi.org/10.1002/nav.3800020109.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.