

1 **Calibration-Free Vehicle Speed Estimation: A Monocular**
2 **Keypoint-Template Approach**

3
4 **Gaofeng Su**

5 Department of Civil & Environmental Engineering
6 The University of California, Berkeley
7 Email: koufongso@berkeley.edu

8 **Keya Li**

9 Department of Civil, Architectural and Environmental Engineering
10 The University of Texas, Austin
11 Email: keya_li@utexas.edu

12 **Raja Sengupta, Ph.D.**

13 Department of Civil & Environmental Engineering
14 The University of California, Berkeley
15 Email: rajasengupta@berkeley.edu

16 **Kara M. Kockelman, Ph.D., P.E***

17 Department of Civil, Architectural and Environmental Engineering
18 The University of Texas, Austin
19 512-471-0210
20 Email: kkockelm@mail.utexas.edu

21
22 Total Number of Pages: 19

23
24 *Submitted: August 3, 2026*

25 *Corresponding author

1 ABSTRACT

2
3 **Objectives:** This paper proposes a calibration-free framework for estimating vehicle speeds from
4 monocular videos reliably and effectively. The method does not rely on roadway features, camera
5 calibration, or roadway features-based reference objects, addressing a key limitation of existing
6 vision-based approaches.

7
8 **Methods:** The proposed calibration-free framework estimates vehicle speeds using a 36-keypoint
9 vehicle template and a homography matrix updated at each frame, requiring no calibration. A
10 YOLO-based keypoint detection module is trained on diverse datasets, and two estimation strategies
11 are compared: keypoint-only tracking and warped optical flow with dense spatial aggregation.
12 Speed is referred by projecting displacements into metric space using the homography, with
13 validation conducted on over 400 video clips from road-side and over-head datasets, covering
14 speeds from 30 to 100 mph.

15
16 **Findings:** The framework achieves reliable speed estimation on the V13 and BrnoCompSpeed
17 datasets, with the warped optical flow method delivering MAEs of 15.0% and 9.7%, respectively,
18 and 77.9% and 93.1% of estimates within $\pm 20\%$ error. After applying a 10% trim to remove edge-
19 of-frame outliers, performance improves to 11.7% and 7.6% MAE, with within- $\pm 20\%$ accuracy
20 rising to 85.3% and 95.4%.

21
22 **Novelty:** This work introduces a calibration-free speed estimation framework that requires no
23 camera calibration or roadway geometry, using a dynamic homography updated frame-by-frame to
24 adapt to vehicle pose changes and mitigate metric drift.

25
26 **Practical Applications:** The method enables low-cost and efficient speed enforcement using
27 portable devices such as dashcams or smartphones, supplementing existing roadway feature-based
28 systems and supporting citizen-based enforcement programs for traffic safety.

1 INTRODUCTION

2 Estimating vehicle speed accurately is a fundamental topic in Intelligent Transportation Systems
3 (ITS) and plays a critical role in traffic enforcement and improving road safety. Active and automated
4 enforcement of speed limits can lower speeds, crash counts, and injuries (Sadeghi-Bazargani and
5 Saadati 2016; Shin et al. 2009). For example, fixed speed cameras can reduce 47% of fatal and
6 injury crashes on urban principal arterials (Shin et al. 2009).

7 Common speed measurement methods include intrusive technologies (e.g., inductive loop detectors)
8 and non-intrusive technologies (e.g., overhead-mounted and side-fired sensors) (Sangsuwan and
9 Ekpanyapong 2024; Odat et al. 2017; Martin et al. 2003). Intrusive detectors are installed
10 within or across the roadway; for example, inductive loop detectors estimate vehicle speeds by
11 detecting vehicle presence within the loop area through changes in oscillator frequency (Martin
12 et al. 2003). Such technologies are expensive to install and can cause disruptions to traffic flow.
13 For instance, a loop system with 40-cm ducts and three chambers costs approximately \$9,000
14 per approach (Webb 2020). Non-intrusive technologies include active detectors (such as active
15 infrared, microwave radar, and ultrasonic sensors) and passive detectors (such as passive infrared,
16 acoustic, and video image processing systems). Active detectors transmit energy (e.g., radar
17 waves or LiDAR) toward vehicles and calculate speeds based on the returned signal, while passive
18 detectors measure energy emitted by vehicles (Javadi et al. 2019). Active sensors generally provide
19 longer detection ranges and greater robustness, but require external power sources and are more
20 complex. In contrast, passive sensors are simpler and require lower power consumption but are
21 more susceptible to environmental conditions (Hovhannisyan 2021).

22 Recently, vision-based speed estimation methods have emerged as a more cost-effective and scalable
23 alternative (Martínez et al. 2024). Over the past decade, computer vision (CV) technologies have
24 been widely used for vehicle detection and tracking (Zhang et al. 2022; Zuraimi and Zaman 2021;
25 Maity et al. 2021), license plate recognition (Xu et al. 2018; Montazzolli and Jung 2017; Parsa
26 et al. 2025), and speed estimation (Hua et al. 2018; Liu et al. 2020). Most state-of-the-art CV-based
27 speed estimation methods typically rely on camera videos and scaling factors derived from road
28 features to estimate travel distances in a two-dimensional (2D) domain using vanishing points (VP)
29 and bounding boxes (Lu et al. 2017; Kim et al. 2018; Chang et al. 2018; Li et al. 2025). Despite
30 their effectiveness, these approaches often require prior knowledge or strong assumptions about
31 roadway geometry, fixed roadway features, or camera calibration, which limits their scalability and
32 applicability.

33 This work introduces a novel and calibration-free framework for vehicle speed estimation that does
34 not rely on roadway features. The proposed approach iteratively updates the homography matrix
35 derived from at least four non-collinear points in the planar video frame, to estimate vehicle speeds
36 and can be readily applied to both overhead- and side-view camera videos. This framework enables
37 speed estimation under flexible deployment conditions and is adaptable to a wide range of roadway
38 environments and camera configurations.

39 The remainder of this paper is organized as follows. Section II reviews related work on CV-based
40 speed estimation. Section III describes the proposed methodology for vehicle speed estimation.
41 Section IV presents the experimental results on overhead and side-view public datasets. Finally,
42 Section V concludes the paper and discusses future work.

1 RELATED WORK

2 CV techniques to estimate vehicle speeds typically begin with video recordings and scene parameters (such as image-to-world scaling factors) as inputs, then use detection and tracking algorithms to estimate traveled distances in the 2D image domain. Speeds are calculated using real-world distances derived from scale factors and the time intervals between consecutive frames (Gunawan et al. 2019; Abdel-Aziz et al. 2015). One critical component of vision-based speed estimation is camera calibration (Zhang 2021; Weng et al. 1992), which enables accurate mapping of 2D image coordinates to 3D real-world motion. Calibration involves estimating both intrinsic parameters (sensor size, resolution, focal length) and extrinsic parameters (camera position and orientation). Traditional methods rely on reference objects with known dimensions, such as lane markings or width, but these require manual measurement and are often impractical for large-scale deployment.

12 To address this, VP-based calibration has emerged as a widely used approach. VPs occur where parallel lines in the 3D world (e.g., lane boundaries, building edges) appear to converge in the image. VP-based methods are classified into geometry-based and learning-based methods. Geometry-based methods use line intersections, clustering techniques, or Gaussian sphere representations to estimate VPs (Feng et al. 2010; Barinova et al. 2010; Collins and Weiss 1990). For instance, Song et al. (2021) adopted an improved Direct Linear Transform (DLT) method by obtaining two VPs from vehicle trajectory lines and car body edge lines, combined with a third VP derived from geometric relationships, to calculate speeds. Results show that the error of the proposed method under different conditions was less than 2.5%. Learning-based methods, instead, use deep neural networks trained on large-scale datasets with VP annotations to infer VPs directly from image context (Zhai et al. 2016; Chang et al. 2018). Hua et al. (2018) combined deep learning-based object detection with optical flow tracking on the NVIDIA AI City Challenge dataset, which contains over 20 hours of traffic videos from multiple camera angles. Their best results achieved a detection rate score of 1.0 and a root mean square error (RMSE) of 12.109 mph. They also pointed out that performance was particularly poor in locations where camera movement due to wind or bridge vibrations degraded video quality. Rani et al. (2024) proposed a deep learning-based framework combining U-Net segmentation with an LV-YOLO network to detect logistic vehicles and estimated their speeds from highway CCTV footage. Results indicate that the method achieved a 2.63% improvement in speed prediction accuracy over the 1D-CNN speed estimation model.

31 Based on estimated VPs and standard camera placement assumptions (zero skew, overhead), intrinsic and extrinsic parameters can be derived, enabling transformation between camera and real-world coordinates. Notably, most existing methods are applied to fixed traffic cameras and standard roadway features, which makes speed estimation particularly challenging in environments where roadway features is unclear or unavailable.

36 PROPOSED METHODOLOGY

37 This section presents the methodology for estimating vehicle speeds from monocular video, based on semantic keypoints, optical flow, and adaptive homography-based mapping. The following subsections outline the problem statement, describe the dynamic homography estimation procedure, derive the speed estimation formulation, analyze errors, and present the complete system pipeline.

1 Problem Statement

2 Given two consecutive images I_k and I_{k+1} capturing N moving vehicles, the objective is to estimate
 3 the set of linear speeds $\mathcal{V}_k = \{v_{i,k}\}_{i=1}^N$, where $v_{i,k}$ denotes the instantaneous speed for vehicle
 4 instance i at time step k . A fundamental challenge in vision-based speed estimation arises from
 5 perspective effects, whereby identical pixel displacements do not correspond to equal physical
 6 distances in the scene. To resolve this, we utilized homography, a projective transformation that
 7 maps points between two planar coordinate frames (Hartley and Zisserman 2003). A key advantage
 8 of this technique is that it does not require prior camera calibration, provided that at least four non-
 9 collinear correspondences between the coordinate frames are identified (Song et al. 2021).

10 Vehicles can be approximated as rigid bodies composed of locally planar surfaces, as illustrated
 11 in Figure 1. The red and yellow shaded regions highlight the distinct lateral and upper planar
 12 faces modeled, respectively, with an independent coordinate frame established for each plane.
 13 Furthermore, it is assumed that the vehicle undergoes longitudinal motion only, with negligible
 14 lateral displacement and rotation between consecutive frames. Under this assumption, the motion
 15 of each planar surface can be represented by a 2D homography.



FIGURE 1: Piecewise planar approximation of a vehicle body. (Frame sampled and modified from the VS13 dataset (Djukanović et al. 2022)).

16 Notation and Preliminaries

17 Let $i \in \{1, \dots, N\}$ denote a specific tracked vehicle instance and k denote the discrete time step.
 18 Each vehicle is modeled as a collection of planar facets $c \in C$, where $C = \{\text{top, left, right}\}$.
 19 Although this work is targeting a multi-vehicle environment, their speeds can be estimated inde-
 20 pendently through an identical process. Moreover, since the homography transformation applies
 21 independently to facets, the subsequent derivation focuses on a single lateral plane of a single
 22 vehicle without loss of generality and maintain a clean and understandable notation.

23 For a tracked vehicle i at time step k , we define:

- 24 1) I_k : The image coordinate frame at time k .
- 25 2) V_k : The metric planar coordinate frame attached to a vehicle facet, and is established based
 26 on the vehicle's pose at time k .
- 27 3) $\mathbf{p}_k \in I_k$: The pixel coordinates $[u, v]^\top$ of a physical point captured at time k .
- 28 4) $\mathbf{s}_k^{V_k}$: The metric coordinates $[x, y]^\top$ of a point captured at time k , expressed in the metric

- 1 frame V_k .
- 2 5) $\mathcal{M}_k$: The set of detected semantic keypoints belonging to a facet at time k .
- 3 6) $\mathbf{F}_k$: The dense optical flow field, where $\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{F}_k(\mathbf{p}_k)$ represents the correspondence
- 4 of the same physical point across consecutive image frames.
- 5 7) v_k : The final estimated scalar speed of the vehicle at time k .

6 Dynamic Image-Template Homography Estimation

7 The projective relationship between frame I_k and V_k is modeled by a homography matrix $\mathbf{H}_k \in \mathbb{R}^{3 \times 3}$.
 8 Using homogeneous coordinates, the mapping from image space to the metric vehicle plane is given
 9 by:

$$10 \quad w \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{H}_k \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (1)$$

11 or equivalently,

$$12 \quad w \tilde{\mathbf{s}}_k^{V_k} = \mathbf{H}_k \tilde{\mathbf{p}}_k \quad (2)$$

13 where $w \neq 0$ is a projective scale factor. The corresponding Euclidean coordinates on the vehicle
 14 plane are recovered by normalizing the homogeneous representation with respect to the third
 15 component, and we expressed it as:

$$16 \quad \mathbf{s}_k^{V_k} = \sigma(\mathbf{H}_k \tilde{\mathbf{p}}_k) \quad (3)$$

17 The homography matrix $\mathbf{H}_k$ can be estimated from at least four non-collinear correspondences
 18 between the semantics keypoint $p_k \in \mathcal{M}_k$, and their corresponding locations on a physical vehicle
 19 metric template. When more than four correspondences are available, $\mathbf{H}_k$ is estimated by solving an
 20 overdetermined system, with Random Sample Consensus (RANSAC) algorithm to reject outliers
 21 (Fischler and Bolles 1981). The mathematical formulation of homography estimation is well
 22 established and is therefore omitted here (Hartley and Zisserman 2003).

23 We defined a vehicle metric template represented by a 2D mechanical drawing labeled with 36
 24 semantic keypoints. Each keypoint is associated with 2D coordinates in meters. Figure 4 illustrates
 25 the sedan-based metric template used in our experiments. To preserve the planar assumption
 26 underlying Equation (1), we selected subsets of approximately co-planar keypoints corresponding
 27 to a specific vehicle facet, for example, wheels center, doors, door handle, side window corners.
 28 By associating detected image keypoints $\mathbf{p}_k$ with their corresponding metric template coordinates
 29 $\mathbf{s}_k^{V_k}$, an instantaneous homography matrix $\mathbf{H}_k$ for each frame k was estimated. Because the metric
 30 scale is enforced by known dimensions of the canonical template, $\mathbf{H}_k$ provides an image-to-metric
 31 mapping that adapts to the vehicle's current position within the field of view.

32 While homography-based methods can theoretically accommodate viewpoint changes, their accu-
 33 racy relies on the assumption that the observed motion on a vehicle surface can be represented
 34 by a single rigid plane. In practice, vehicle surfaces are only approximately planar, and vehicle
 35 motion introduces deviations due to suspension dynamics, minor rotations, and depth variations
 36 across keypoints. These effects cause the true image-to-plane transformation to gradually deviate

from the estimated homography, resulting in persistent metric errors. Moreover, the homography is estimated from keypoint correspondences at a specific vehicle pose and is therefore only valid for that reference configuration. As the vehicle moves, the tracked keypoints may undergo increasingly large image displacements relative to their original positions, while the fixed homography cannot account for the resulting changes in projection. Consequently, planar approximation errors and temporal homography mismatch accumulate over time, leading to drift in displacement and speed estimates.

To address this limitation, this paper models the homography as a dynamic, time-varying transformation. By re-estimating the homography at each time step using keypoints associated with the same vehicle planar facet, the proposed approach continuously adapts the planar approximation to the vehicle's current pose. This dynamic estimation reduces the accumulation of projection bias and improves the consistency of displacement and speed estimates over time.

Speed Estimation via Warped Optical Flow

At time step k , a planar metric coordinate frame V_k is defined for each tracked vehicle facet. Given a homography $\mathbf{H}_k$ between the image frame I_k and the metric planar frame V_k , the metric displacement of an arbitrary point on the facet can be recovered if its image correspondence between consecutive frames is known. Specifically, for a point $\mathbf{p}_{i,k}$ and its corresponding location $\mathbf{p}_{i,k+1}$ at the next time step, the instantaneous speed is computed as

$$v_{s_{i,k}} = \frac{\|\sigma(\mathbf{H}_k \tilde{\mathbf{p}}_{i,k+1}) - \sigma(\mathbf{H}_k \tilde{\mathbf{p}}_{i,k})\|}{\Delta t}, \quad (4)$$

where $\tilde{\mathbf{p}}$ represents the point in homogeneous coordinates, $\sigma(\cdot)$ denotes the homogeneous normalization operator, and $v_{s_{i,k}}$ represents the metric speed of the tracked point. Note that $\mathbf{H}_k$ is only valid for points belonging to the same planar facet used during its estimation.

The required point correspondence can be obtained using either dense or sparse tracking strategies. For dense correspondence, the optical flow field $\mathbf{F}_k$ is used to obtain the subsequent image location:

$$\mathbf{p}_{i,k+1} = \mathbf{p}_{i,k} + \mathbf{F}_k(\mathbf{p}_{i,k}). \quad (5)$$

Alternatively, sparse correspondence can be obtained by detecting and matching semantic keypoints that appear in both consecutive frames. This provides direct correspondences between $\mathbf{p}_{i,k}$ and $\mathbf{p}_{i,k+1}$ without relying on dense optical flow estimation.

For the optical flow strategy, a convex hull $\mathcal{S}_k$ is constructed from the detected semantic keypoints $\mathbf{p}_k \in \mathcal{M}_k$ to restrict the queried points to the same planar facet. Optical flow is then evaluated only for points within this convex hull, ensuring that the resulting correspondences remain valid under the facet-specific homography $\mathbf{H}_k$. Moreover, this constraint increases the number of available tracking points beyond the sparse semantic keypoints, enabling robust aggregation over multiple measurements and reducing the influence of individual optical flow outliers.

The final vehicle speed estimate is obtained by aggregating the point-wise speed estimates over all tracked points in $\mathcal{S}_k$:

$$v_k = \text{RobustAgg}(\{v_{s_{i,k}} \mid s_{i,k} \in \mathcal{S}_k\}), \quad (6)$$

where $\text{RobustAgg}(\cdot)$ denotes a robust aggregation function used to reduce the influence of optical flow outliers and keypoint localization errors.

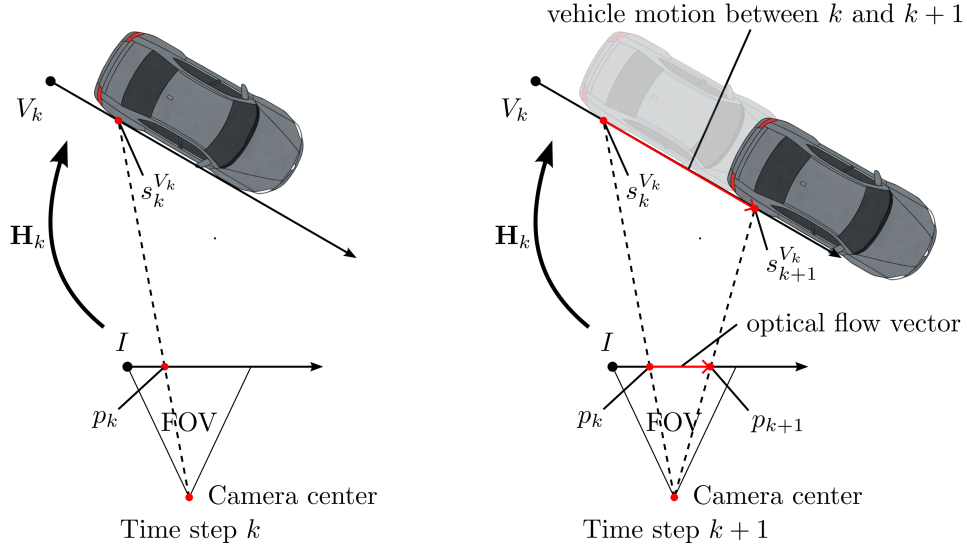


FIGURE 2: Illustration of speed estimation via warped optical flow for a tracked vehicle's right facet. The vehicle graphic was generated using AI tools.

4 Error Analysis

The speed estimate in (4) depends on the accuracy of the estimated homography matrix, which is recovered from image keypoint correspondences, and the accuracy of the optical flow. To quantify this effect, we analyzed the first-order sensitivity of the proposed speed estimation method to homography estimation errors.

For notation simplicity, the subscript k is dropped in the following analysis. Let H^* denote the true homography matrix, and ΔH denote the homography matrix estimation error, yielding:

$$H = H^* + \Delta H \quad (7)$$

Similarly, let $s_1 = s_0 + v_{OF}^*$ represents the true correspondence between two frames, where v_{OF}^* is the true optical flow vector. Accounting for optical flow error Δv_{OF} , we have:

$$v_{OF} = v_{OF}^* + \Delta v_{OF} \quad (8)$$

Using Equation (4), the true velocity v^* satisfies:

$$v^* \Delta t = \sigma(H^* s_1^*) - \sigma(H^* s_0^*) \quad (9)$$

1 While the estimated velocity v is computed as:

$$2 \quad v = \Delta t^{-1}(\sigma(Hs_1) - \sigma(Hs_0)) \quad (10)$$

$$3 \quad = \Delta t^{-1}(\sigma((H^* + \Delta H)s_1) - \sigma((H^* + \Delta H)s_0)) \quad (11)$$

4 The first-order approximation of the normalization function σ gives:

$$5 \quad \sigma((H^* + \Delta H)s) \approx \sigma(H^*s) + J_\sigma(H^*s)\Delta Hs \quad (12)$$

6 where J_σ is the Jacobian of the normalization operator σ . Consequently,

$$7 \quad v - v^* \approx \Delta t^{-1}(J_\sigma(H^*s_1)\Delta Hs_1 - J_\sigma(H^*s_0)\Delta Hs_0) \quad (13)$$

8 Assuming small inter-frame motion such that $J_\sigma(H^*s_1) \approx J_\sigma(H^*s_0) = J$, it follows that:

$$9 \quad v - v^* \approx \Delta t^{-1}J\Delta H(s_1 - s_0) = \Delta t^{-1}J\Delta Hv_{OF} \quad (14)$$

10 The error is bounded by:

$$11 \quad \|\Delta v\| = \|v - v^*\| \leq \Delta t^{-1}\|J\|\|\Delta H\|\|v_{OF}\| \quad (15)$$

12 The bound in (15) indicates that the speed estimation error is affected by three components: the
 13 perspective sensitivity $\|J_\sigma\|$, the homography estimation error $\|\Delta H\|$, and the magnitude of the
 14 optical flow $\|v_{OF}\|$. The first term reflects geometric amplification driven by the viewing angle
 15 between the vehicle facet and the image plane. The second term depends on the conditioning of the
 16 linear system used to solve for the homography. The third term shows that larger image motions
 17 linearly scale the propagation of homography errors into the final metric velocity.

18 Integrated System Pipeline

19 Building upon the theoretical framework established in the previous section, the architecture of the
 20 proposed integrated system is illustrated in Figure 3. The pipeline processes a sequence of video
 21 frames to simultaneously extract dense motion features and semantic vehicle attributes.

22 For each incoming frame pair $\{I_k, I_{k+1}\}$, a dense optical flow field $\mathbf{F}_k$ is computed to capture pixel-
 23 level displacements. In parallel, a Multi-Object Tracking (MOT) module identifies, localizes and
 24 generate bounding boxes for each vehicle. Based on these bounding boxes, the image is cropped to
 25 generate vehicle-specific Regions of Interest (ROIs). This spatial isolation ensures that keypoints
 26 are detected exclusively within the context of the target vehicle, thereby minimizing background
 27 noise and enhancing the accuracy of the detection model.

28 The core component of the proposed system is the *Estimate Speed and Update Homography* module,
 29 which integrates semantic keypoints, optical flow, and adaptive homography estimation for metric
 30 speed recovery. At each iteration, the module executes a three-step procedure:

- 31 1) **Keypoint Detection and Classification:** First, a specialized keypoint detection network iden-
 32 tifies semantic keypoints within the vehicle ROI and classifies them according to predefined
 33 vehicle facets (e.g., the left wheel center keypoint belongs to the left facet).
- 34 2) **Speed Estimation:** For each facet, the system retrieves the associated keypoints and con-

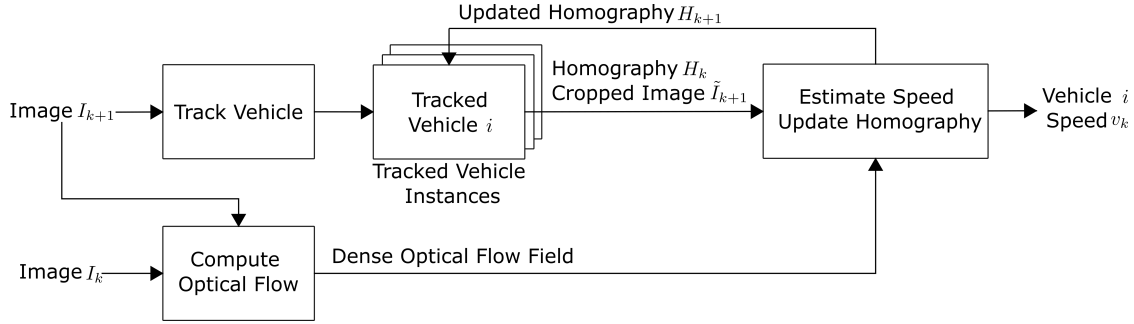


FIGURE 3: Proposed System Architecture.

1 constructs the convex hull $\mathcal{S}_k$, employing the homography matrix $\mathbf{H}_k$ established at the previous
 2 time step. Using the inter-frame time interval Δt and the optical flow vectors, the metric
 3 speed for each point $s_{i,k} \in \mathcal{S}_k$ is calculated via Equation 4. To determine the final vehicle
 4 speed, we aggregate the individual point speeds using an arithmetic mean. Note that more
 5 robust aggregation functions (e.g., RANSAC or median filtering) may be employed here to
 6 mitigate the influence of outliers.

- 7 3) **Geometric Calibration (Homography Update):** Finally, the homography matrix is updated
 8 for the next iteration. For any facet containing a sufficient number of valid keypoints (specif-
 9 ically, $N \geq 4$ non-collinear points), the homography is recomputed to account for changes in
 10 the vehicle's perspective or road geometry.

11 This recursive estimate-then-update strategy ensures that the metric scaling remains tightly coupled
 12 to the vehicle's evolving perspective and distance. By continuously refining the homography
 13 matrices based on new keypoint detections, the system effectively resets accumulated tracking
 14 error for every frame in the sequence, mitigating drift over long tracking durations.

15 It is important to note that this design introduces a deterministic one-frame latency, as the speed
 16 estimate for time step k relies on the optical flow derived from the arrival of the subsequent frame
 17 I_{k+1} . While technically acausal with respect to the instantaneous frame I_k , this forward-looking
 18 dependency is inherent to dense optical flow algorithms, which require temporal displacement
 19 to compute motion. However, given the high temporal resolution of modern camera hardware,
 20 typically ≥ 30 frames per second (FPS), this results in a negligible latency of approximately 33
 21 ms. This delay is well within the acceptable operational bounds for real-time applications such as
 22 traffic flow analysis and automated speed enforcement.

23 EXPERIMENT RESULT AND DISCUSSION

24 Experiments were conducted to evaluate the effectiveness of the proposed vehicle speed estimation
 25 framework. The evaluation focuses on the accuracy of the recovered vehicle speeds, the robustness
 26 of the proposed homography-based metric mapping, and the performance of the system under
 27 varying vehicle poses and camera viewpoints.

28 A sedan model consisting of 36 semantic keypoints was used as the metric template, as shown in
 29 Figure 4. Each keypoint was assigned a 2D (image) coordinates in meters, assuming a wheelbase
 30 (horizontal) distance of 3 meters. The keypoint detection module was trained using the Ultralytics

1 YOLO framework with the YOLO-11s backbone (Jocher et al. 2026). The training dataset
 2 consists of 756 images for training and 247 images for validation, collected from the Waymo Open
 3 Dataset (Sun et al. 2020), Roundabout HD dataset (Lin et al. 2025), and BoxCar dataset (Sochor
 4 et al. 2018a).

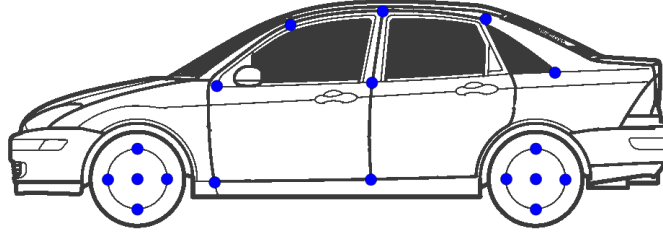


FIGURE 4: 36 semantic keypoints for the sedan are marked with blue circles (symmetric on the other side). Assuming a wheelbase of approximately 3 meters. Car graphics were adapted from Wikipedia. (842U 2010)

5 The validation experiments were conducted using video sequences from the VS13 (resolution 1920
 6 $\times 1080$, 30FPS) (Djukanović et al. 2022) and BrnoCompSpeed datasets (resolution 1920×1080 ,
 7 50FPS) (Sochor et al. 2018b), with total 443 video clips across 208 unique vehicles and speed
 8 ranging from 30 mph to 100 mph. None of these evaluation sequences were part of the keypoint
 9 detection model’s training data. The two datasets offer two different camera viewpoints: VS13
 10 provides pedestrian-level angles of approaching traffic, while BrnoCompSpeed provides overhead
 11 surveillance views. This allows us to test the framework across different viewing geometries and
 12 motion profiles. Figure 5 shows sample processed frames from each dataset, with tracked cars
 13 marked by green bounding boxes and detected semantic keypoints highlighted as blue circles. We
 14 also compare two distinct speed estimation strategies: a Keypoint-only (KP) approach that tracks
 15 sparse semantic features, and our proposed Warped Optical Flow (OF) method, which leverages a
 16 keypoint-defined convex hull to achieve robust, dense spatial aggregation. All errors are presented
 17 as relative percentages for comparison across different speeds.



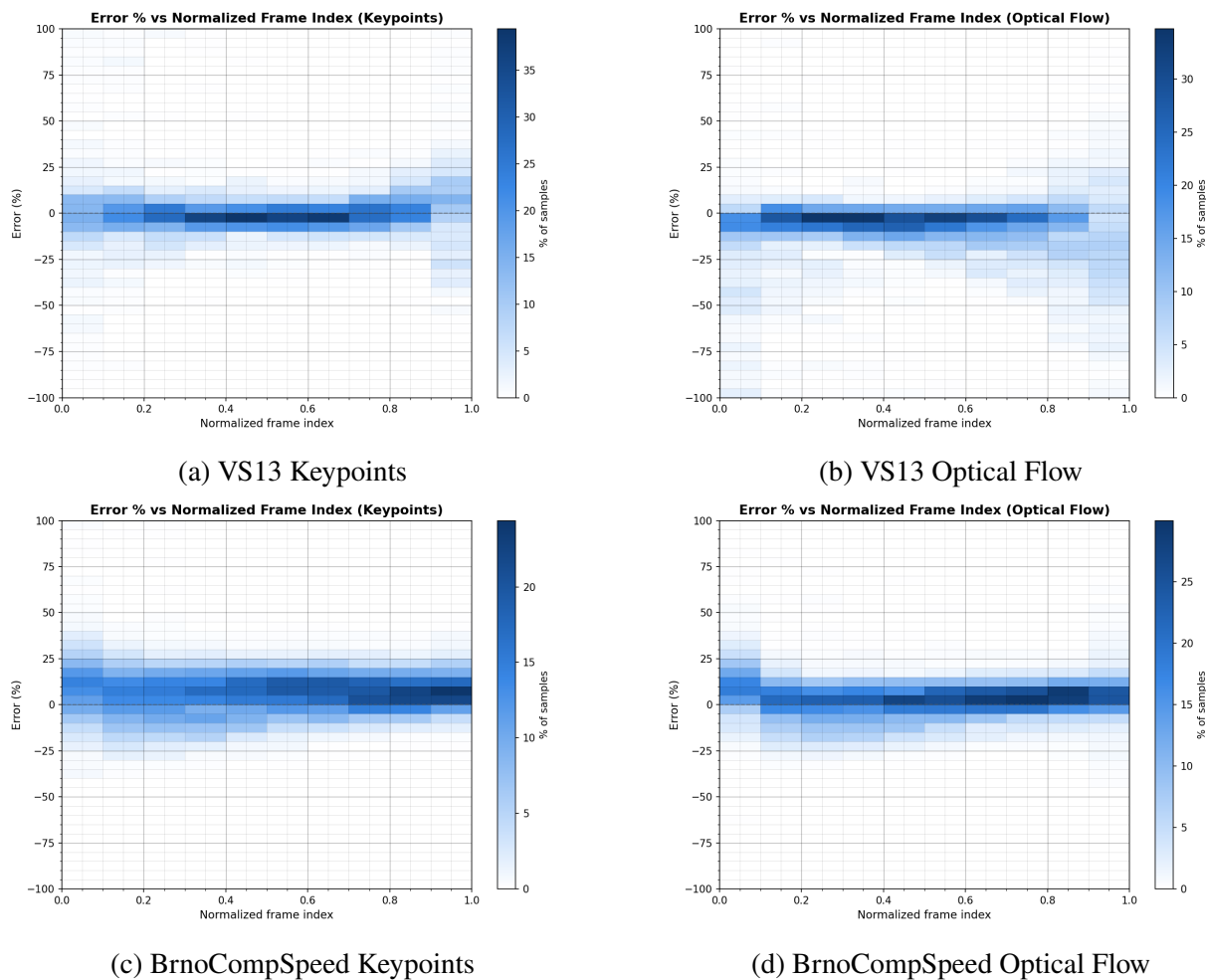
FIGURE 5: Sample image from the VS13 dataset (left) and BrnoCompSpeed dataset (right), annotated using the speed estimation framework.

18 Table 1 presents the overall accuracy of speed estimation in both datasets. For the VS13 dataset,
 19 the optical flow method achieves a Mean Absolute Error (MAE) of 15.0% with a median error
 20 of -6.0% and 77.9% of estimates within $\pm 20\%$ error, compared to the keypoint method which

TABLE 1: Summary of speed estimation accuracy for VS13 and BrnoCompSpeed datasets.

Data	Method	Group	N	Med.	MAE	RMSE	Thresholds			
							$\pm 5\%$	$\pm 10\%$	$\pm 20\%$	$> 50\%$
VS13	KP	All	4,317	-0.9%	20.8%	178.4%	46.7%	72.5%	87.6%	4.4%
		Sedan	2,055	-0.7%	24.3%	243.0%	46.0%	73.3%	89.7%	3.2%
		NoSed	2,262	-1.0%	17.7%	84.2%	47.3%	71.7%	85.7%	5.6%
	OF	All	4,578	-6.0%	15.0%	53.5%	37.3%	58.8%	77.9%	5.3%
		Sedan	2,195	-6.1%	14.3%	22.1%	35.4%	56.8%	76.1%	4.8%
		NoSed	2,383	-6.0%	15.6%	71.1%	39.1%	60.7%	79.6%	5.7%
BC	KP	All	17,085	6.5%	12.8%	46.9%	28.3%	54.3%	87.9%	1.2%
		Sedan	3,361	6.9%	11.1%	28.8%	32.0%	58.9%	89.0%	0.6%
		NoSed	13,724	6.3%	13.2%	50.4%	27.3%	53.1%	87.6%	1.3%
	OF	All	17,600	3.0%	9.7%	66.1%	41.4%	70.6%	93.1%	1.0%
		Sedan	3,452	3.9%	11.0%	141.0%	47.5%	75.1%	93.9%	0.7%
		NoSed	14,148	2.7%	9.3%	24.2%	39.9%	69.5%	92.9%	1.0%

Note: KP = Keypoints; OF = Optical Flow. Med. = Median; MAE = Mean Absolute Error; RMSE = Root Mean Square Error; N = Number of frames with valid estimation.

**FIGURE 6:** Error Percentage vs. Frame Index for VS13 and BrnoCompSpeed datasets

achieves an MAE of 20.8% with a median error of -0.9% and 87.6% within $\pm 20\%$. Although the keypoint method achieves higher within-threshold accuracy, which is particularly relevant for speed limit enforcement, the optical flow method yields significantly lower MAE and RMSE, indicating more consistent and accurate estimates overall. For the BrnoCompSpeed dataset, both methods perform significantly better, with optical flow achieving an MAE of 9.7% and 93.1% within $\pm 20\%$, compared to 12.8% MAE and 87.9% within $\pm 20\%$ for the keypoint method. The optical flow method consistently outperforms the keypoint approach across all metrics on this dataset. Across both datasets, optical flow demonstrates higher within-threshold accuracy. Performance differences between sedans and non-sedans are very small, with sedans generally achieving slightly lower MAE with optical flow, likely due to the dataset being more trained on sedans.

To investigate the temporal behavior of estimation errors, the error percentage was analyzed as a function of the normalized frame index. Specifically, frame indices were scaled to a range of 0 to 1 based on the first and last frames containing a valid speed estimation. Figure 6 displays heatmaps of error percentage versus normalized frame index for both methods and datasets. The heatmaps indicate that errors were significantly higher at the beginning and the end of each video sequence, when vehicles were either too far from the camera—resulting in a small pixel footprint—or too close, partially leaving the frame and making feature tracking unreliable. This directly increased homography estimation error.

Furthermore, the OF method became particularly unstable at extreme vehicle-to-camera distances. This vulnerability is more obvious in the VS13 scenario where vehicles move directly toward the camera, resulting in dramatic scale changes. Consequently, OF exhibits a much wider error spread in Figure 6(b) compared to BrnoCompSpeed, where the camera is positioned overhead and the relative distance change remains minimal across the frames.

To mitigate the effects of unreliable detections at the very first and last frame, a 10% trim was applied to each tail of the error distribution. Table 2 summarizes the results after this trimming. For the VS13 dataset, the keypoint method's MAE improved from 20.8% to 14.3%, while its within $\pm 20\%$ accuracy increased from 87.6% to 94.0%. The optical flow method also improved from 15.0% to 11.7% in MAE, with within $\pm 20\%$ rising from 77.9% to 85.3%. In terms of the BrnoCompSpeed dataset, the optical flow method's MAE dropped from 9.7% to 7.6%, and the keypoint method's MAE dropped from 12.8% to 11.4%. The trimming also significantly reduced outlier rates. For example, for the VS13 dataset, the keypoint method's outlier rate decreased from 4.4% to 2.1%, while the optical flow method's outlier rate decreases from 5.3% to 2.3%. Both method also showed enhanced performance on BrnoCompSpeed dataset after trimming.

Figure 7 compares the speed estimation errors across different methods and datasets, with ground truth speeds ranging from 20 to 70 mph. The most accurate estimations occur in the 40 to 60 mph speed range, where both optical flow-based methods achieve errors below 5%. In contrast, estimation errors are largest at low speeds (20 to 30 mph), particularly for keypoint-based methods, where errors exceed 20%, possibly due to increased sensitivity to noise.

CONCLUSION

This paper demonstrates that monocular camera-based vehicle speed estimation can be effectively achieved using a vehicle metric template, a lightweight vehicle keypoint detection module, and

1 a homography technique. Crucially, the proposed method operates without requiring camera
 2 calibration (intrinsic or extrinsic) or pre-existing roadway features, relying only on a relatively
 3 stationary camera view. The approach achieves reliable and accurate speed estimation, with 87.6%
 4 and 87.9% of estimates falling within the $\pm 20\%$ error threshold across the evaluated datasets,
 5 demonstrating its effectiveness under diverse viewpoints and driving conditions.

TABLE 2: Summary of speed estimation accuracy for VS13 and BrnoCompSpeed datasets with 10% trimming.

Data	Method	Group	N	Med.	MAE	RMSE	Thresholds			
							$\pm 5\%$	$\pm 10\%$	$\pm 20\%$	$> 50\%$
VS13	KP	All	3,334	-1.2%	14.3%	170.6%	54.0%	81.0%	94.0%	2.1%
		Sedan	1,584	-1.2%	18.5%	240.9%	52.0%	79.9%	94.1%	1.8%
		NoSed	1,750	-1.3%	10.5%	54.1%	55.7%	82.0%	93.9%	2.3%
	OF	All	3,555	-5.1%	11.7%	57.4%	43.0%	66.2%	85.3%	2.3%
		Sedan	1,692	-5.4%	11.6%	17.7%	40.1%	62.4%	81.7%	2.4%
		NoSed	1,863	-4.8%	11.7%	77.4%	45.6%	69.7%	88.5%	2.1%
BC	KP	All	13,822	6.0%	11.4%	28.4%	28.5%	55.0%	89.4%	0.6%
		Sedan	2,738	6.8%	9.5%	11.9%	32.8%	60.3%	90.8%	0.0%
		NoSed	11,084	5.7%	11.8%	31.2%	27.4%	53.6%	89.1%	0.8%
	OF	All	14,198	2.3%	7.6%	11.4%	43.9%	73.9%	95.4%	0.5%
		Sedan	2,794	3.4%	6.5%	8.9%	50.9%	79.1%	96.7%	0.1%
		NoSed	11,404	1.9%	7.9%	12.0%	42.2%	72.7%	95.1%	0.6%

Note: KP = Keypoints; OF = Optical Flow. Med. = Median; MAE = Mean Absolute Error; RMSE = Root Mean

Square Error; N = Number of frames with valid estimation.

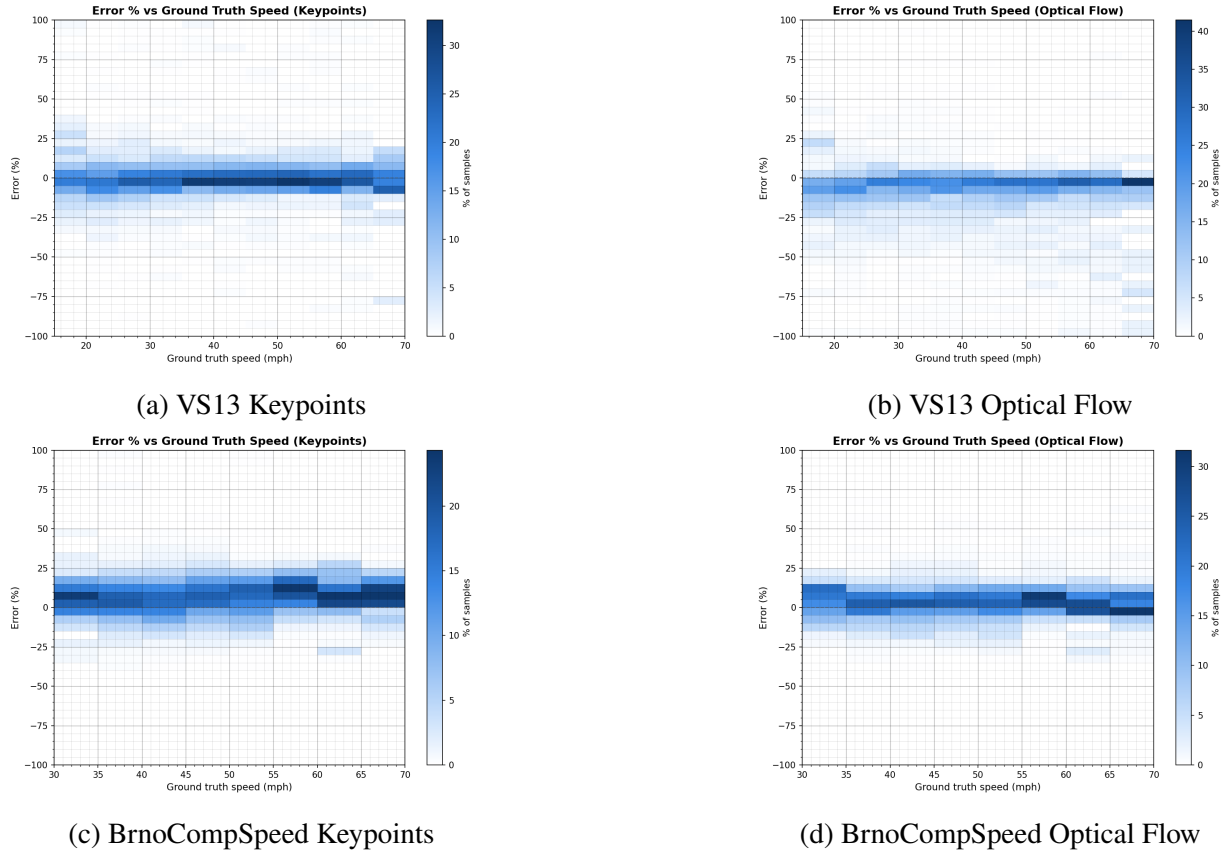


FIGURE 7: Error Percentage vs. Speed for VS13 and BrnoCompSpeed datasets

1 Experimental results also indicate that correspondence computation via the proposed warped
 2 optical flow method is more reliable and robust against outliers. In contrast, purely keypoint-based
 3 correspondence can offer high peak accuracy but remains sensitive to keypoint detection jitter.

4 Additionally, this work has significant practical implications for traffic enforcement and public
 5 safety. Unlike traditional speed measurement systems that rely on roadway features, the proposed
 6 method is deployable and cost-effective. This facilitates citizen-based enforcement, where the
 7 public can use portable devices, such as dashcams or smartphones, to document speeding events
 8 (Li et al. 2025). Nowadays, many cities have piloted citizen-based enforcement programs. For
 9 example, New York City has empowered citizens to enforce diesel-truck idling laws by submitting
 10 video evidence; those who provide at least three minutes of footage receive 25% of any fine collected,
 11 amounting to about \$87.50 per violation (Hollis 2026). As vision-based enforcement becomes
 12 increasingly popular due to budget constraints and the rapid growth of camera-equipped devices,
 13 the proposed approach delivers a solution that can complement existing enforcement mechanisms
 14 while reducing the cost of camera installation.

15 However, the current implementation has certain limitations. It relies on a generalized sedan metric
 16 template, which can introduce minor errors due to keypoint detection inaccuracies and homography
 17 transformations. Although experiments show that the framework maintains decent accuracy on non-
 18 sedan vehicles, future work will focus on expanding the template library to include diverse vehicle
 19 classes and geometries. Additionally, while the warped optical flow approach mitigates individual

1 outliers, achieving even higher accuracy will likely require more robust and reliable correspondence
2 tracking methods to handle challenging conditions such as illumination changes or vehicles in close
3 proximity to the camera.

4 **ACKNOWLEDGMENTS**

5 The authors thank Helena Chandy for her valuable insights and edits.

6 **DECLARATION OF CONFLICTING INTERESTS**

7 All authors declare no potential conflicts of interest with respect to the research, authorship, and
8 publication of this article.

1 REFERENCES

- 2 Sadeghi-Bazargani, H. and M. Saadati, Speed management strategies; a systematic review. *Bulletin*
- 3 *of Emergency & Trauma*, Vol. 4, No. 3, 2016, p. 126.
- 4 Shin, K., S. P. Washington, and I. van Schalkwyk, Evaluation of the Scottsdale Loop 101 automated
- 5 speed enforcement demonstration program. *Accident Analysis & Prevention*, Vol. 41, No. 3,
- 6 2009, pp. 393–403.
- 7 Sangsuwan, K. and M. Ekpanyapong, Video-based vehicle speed estimation using speed measure-
- 8 ment metrics. *IEEE Access*, Vol. 12, 2024, pp. 4845–4858.
- 9 Odat, E., J. S. Shamma, and C. Claudel, Vehicle classification and speed estimation using combined
- 10 passive infrared/ultrasonic sensors. *IEEE transactions on intelligent transportation systems*,
- 11 Vol. 19, No. 5, 2017, pp. 1593–1606.
- 12 Martin, P. T., Y. Feng, X. Wang, et al., *Detector technology evaluation*. Mountain-Plains Consortium
- 13 Fargo, ND, USA, 2003.
- 14 Webb, C., *The Economics of Inductive Loops vs. Radar Traffic Detection*. [https://ct-technologyinf](https://ct-technologyinfo.com/blog/2020/11/09/traffic-detection-systems/)
- 15 [o.com/blog/2020/11/09/traffic-detection-systems/](https://ct-technologyinfo.com/blog/2020/11/09/traffic-detection-systems/), 2020, Accessed: 2026-01-04.
- 16 Javadi, S., M. Dahl, and M. I. Pettersson, Vehicle speed measurement model for video-based
- 17 systems. *Computers & electrical engineering*, Vol. 76, 2019, pp. 238–248.
- 18 Hovhannisyan, T., *Active vs Passive sensing: Is one better than the other?* [https://www.rvmagnet](https://www.rvmagnetics.com/active-vs-passive-sensing-is-one-better-than-the-other-116)
- 19 [ics.com/active-vs-passive-sensing-is-one-better-than-the-other-116](https://www.rvmagnetics.com/active-vs-passive-sensing-is-one-better-than-the-other-116), 2021, Updated: Aug.
- 20 2, 2023; Jul. 22, 2024. Accessed: 2026-01-05.
- 21 Martínez, A. H., I. G. Daza, C. F. López, and D. F. Llorca, Digital twins to alleviate the need for real
- 22 field data in vision-based vehicle speed detection systems. In *2024 IEEE 27th International*
- 23 *Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2024, pp. 2536–2541.
- 24 Zhang, Y., Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, Real-time vehicle detection based on
- 25 improved yolo v5. *Sustainability*, Vol. 14, No. 19, 2022, p. 12274.
- 26 Zuraimi, M. A. B. and F. H. K. Zaman, Vehicle detection and tracking using YOLO and DeepSORT.
- 27 In *2021 IEEE 11th IEEE symposium on computer applications & industrial electronics*
- 28 *(ISCAIE)*, IEEE, 2021, pp. 23–29.
- 29 Maity, M., S. Banerjee, and S. S. Chaudhuri, Faster r-cnn and yolo based vehicle detection: A survey.
- 30 In *2021 5th international conference on computing methodologies and communication*
- 31 *(ICCMC)*, IEEE, 2021, pp. 1442–1447.
- 32 Xu, Z., W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, and L. Huang, Towards end-to-end
- 33 license plate detection and recognition: A large dataset and baseline. In *Proceedings of the*
- 34 *European conference on computer vision (ECCV)*, 2018, pp. 255–271.
- 35 Montazzolli, S. and C. Jung, Real-time brazilian license plate detection and recognition using deep
- 36 convolutional neural networks. In *2017 30th SIBGRAPI conference on graphics, patterns*
- 37 *and images (SIBGRAPI)*, IEEE, 2017, pp. 55–62.
- 38 Parsa, P., K. Li, K. M. Kockelman, and S. Choi, Video-based Vehicle Surveillance in the Wild:
- 39 License Plate, Make, and Model Recognition with Self Reflective Vision-Language Models.
- 40 *arXiv preprint arXiv:2508.01387*, 2025.
- 41 Hua, S., M. Kapoor, and D. C. Anastasiu, Vehicle tracking and speed estimation from traffic
- 42 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*
- 43 *Workshops*, 2018, pp. 153–160.

- 1 Liu, C., D. Q. Huynh, Y. Sun, M. Reynolds, and S. Atkinson, A vision-based pipeline for vehicle
2 counting, speed estimation, and classification. *IEEE transactions on intelligent transporta-*
3 *tion systems*, Vol. 22, No. 12, 2020, pp. 7547–7560.
- 4 Lu, X., J. Yaoy, H. Li, Y. Liu, and X. Zhang, 2-line exhaustive searching for real-time vanishing
5 point estimation in manhattan world. In *2017 IEEE Winter Conference on Applications of*
6 *Computer Vision (WACV)*, IEEE, 2017, pp. 345–353.
- 7 Kim, J.-H., W.-T. Oh, J.-H. Choi, and J.-C. Park, Reliability verification of vehicle speed estimate
8 method in forensic videos. *Forensic science international*, Vol. 287, 2018, pp. 195–206.
- 9 Chang, C.-K., J. Zhao, and L. Itti, Deepvp: Deep learning for vanishing point detection on 1 million
10 street view images. In *2018 IEEE International Conference on Robotics and Automation*
11 *(ICRA)*, IEEE, 2018, pp. 4496–4503.
- 12 Li, K., J. Malagavalli, L. Goel, T. Wang, and K. Kockelman, Smartphone-based Method for
13 Automated Speed Enforcement. In *2025 Transportation Research Board Annual Meeting,*
14 *National Academy of Sciences*, 2025.
- 15 Gunawan, A. A., D. A. Tanjung, and F. E. Gunawan, Detection of vehicle position and speed using
16 camera calibration and image projection methods. *Procedia Computer Science*, Vol. 157,
17 2019, pp. 255–265.
- 18 Abdel-Aziz, Y. I., H. M. Karara, and M. Hauck, Direct linear transformation from comparator
19 coordinates into object space coordinates in close-range photogrammetry. *Photogrammetric*
20 *engineering & remote sensing*, Vol. 81, No. 2, 2015, pp. 103–107.
- 21 Zhang, Z., Camera calibration. In *Computer vision: a reference guide*, Springer, 2021, pp. 130–131.
- 22 Weng, J., P. Cohen, M. Herniou, et al., Camera calibration with distortion models and accuracy
23 evaluation. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 14, No. 10,
24 1992, pp. 965–980.
- 25 Feng, C., F. Deng, and V. R. Kamat, Semi-automatic 3d reconstruction of piecewise planar building
26 models from single image. *CONVR (Sendai:)*, Vol. 2, No. 5, 2010, p. 6.
- 27 Barinova, O., V. Lempitsky, E. Tretyak, and P. Kohli, Geometric image parsing in man-made
28 environments. In *European conference on computer vision*, Springer, 2010, pp. 57–70.
- 29 Collins, R. T. and R. S. Weiss, Vanishing point calculation as a statistical inference on the unit
30 sphere. In *ICCV*, 1990, Vol. 90, pp. 400–403.
- 31 Song, J., H. Song, and S. Wang, PTZ camera calibration based on improved DLT transformation
32 model and vanishing Point constraints. *Optik*, Vol. 225, 2021, p. 165875.
- 33 Zhai, M., S. Workman, and N. Jacobs, Detecting vanishing points using global image context in
34 a non-manhattan world. In *Proceedings of the IEEE Conference on Computer Vision and*
35 *Pattern Recognition*, 2016, pp. 5657–5665.
- 36 Rani, N. G., N. H. Priya, A. Ahilan, and N. Muthukumaran, LV-YOLO: logistic vehicle speed
37 detection and counting using deep learning based YOLO network. *Signal, Image and Video*
38 *Processing*, Vol. 18, No. 10, 2024, pp. 7419–7429.
- 39 Hartley, R. and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university
40 press, 2003.
- 41 Djukanović, S., N. Bulatović, and I. Čavor, A dataset for audio-video based vehicle speed estimation.
42 In *2022 30th Telecommunications Forum (TELFOR)*, IEEE, 2022, pp. 1–4.
- 43 Fischler, M. A. and R. C. Bolles, Random sample consensus: a paradigm for model fitting with
44 applications to image analysis and automated cartography. *Communications of the ACM*,
45 Vol. 24, No. 6, 1981, pp. 381–395.

- 1 Jocher, G., J. Qiu, M. Liu, S. Lyu, F. C. Akyon, and M. E. Kalfaoglu, Ultralytics YOLO26: Unified
2 Real-Time End-to-End Vision Models, 2026.
- 3 Sun, P., H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai,
4 B. Caine, et al., Scalability in perception for autonomous driving: Waymo open dataset.
5 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
6 2020, pp. 2446–2454.
- 7 Lin, Y., S. Lockyer, M. Sui, L. Gan, F. Stanek, M. Zarbock, W. Li, A. Evans, and N. Zhang,
8 RoundaboutHD: High-Resolution Real-World Urban Environment Benchmark for Multi-
9 Camera Vehicle Tracking. *arXiv preprint arXiv:2507.08729*, 2025.
- 10 Sochor, J., J. Špaňhel, and A. Herout, Boxcars: Improving fine-grained recognition of vehicles using
11 3-d bounding boxes in traffic surveillance. *IEEE transactions on intelligent transportation*
12 *systems*, Vol. 20, No. 1, 2018a, pp. 97–108.
- 13 842U, *Three body styles with pillars and boxes.png*. *Wikimedia Commons*. [https://commons.](https://commons.wikimedia.org/wiki/File:Three_body_styles_with_pillars_and_boxes.png)
14 [wikimedia.org/wiki/File:Three_body_styles_with_pillars_and_boxes.png](https://commons.wikimedia.org/wiki/File:Three_body_styles_with_pillars_and_boxes.png). Accessed:
15 2026-07-29, 2010.
- 16 Sochor, J., R. Juránek, J. Špaňhel, L. Maršík, A. Široký, A. Herout, and P. Zemčík, Comprehensive
17 data set for automatic single camera visual speed measurement. *IEEE Transactions on*
18 *Intelligent Transportation Systems*, Vol. 20, No. 5, 2018b, pp. 1633–1643.
- 19 Hollis, D., Trucking association sues New York City over idling law, citizen videos. *Truckers News*,
20 2026, updated July 14, 2026.