Review | Received 28 January 2024; Accepted 17 June 2024; Published 11 July 2024
https://doi.org/10.55092/aias20240004

Vehicle speed measurement technologies in Intelligent Transportation Systems: current status, challenges and future directions

Zhili Chen, **Fang Guo*** **and Longmei Luo**

School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou, China

* Correspondence author; E-mail: Davidace@fjut.edu.cn.

Abstract: Speed measurement is essential for the development of Intelligent Traffic Systems (ITS), and the adoption and enforcement of appropriate speed limits are among the most effective strategies to improve road safety. This review offers an exhaustive exploration of vehicle speed measurement methods and technologies within traffic applications. While inductive loop detectors and radar are mature technologies in traffic speed measurement, cameras are typically used to facilitate license plate recognition. This paper delves into the principles and technologies behind traditional speed measurement systems such as inductive loop detectors, wireless radar, LiDAR, and the Global Positioning System, alongside computer vision-based speed measurement. It examines the evolution of computer vision, reviews common datasets, and explores the feasibility of using cameras for direct speed measurement. Furthermore, this paper evaluates the precision, cost, and practicality of these technologies and discusses future research directions, providing crucial references and guidance for advancing Intelligent Traffic Systems.

Keywords: intelligent traffic; vehicle speed measurement; accuracy; multi-sensor fusion; long-range

1. Introduction

In 1993, the Intelligent Transportation Society of America introduced the concept of Intelligent Transportation Systems (ITS) at the IVHS World Congress [1]. ITS leverages advanced information technology, communication technology, and control technology to enhance the efficiency, safety, sustainability, and convenience of transportation systems. It emphasizes the critical role of real-time vehicle speed measurement as a key function of ITS. With technological progress, vehicle speed measurement has rapidly evolved, transitioning from traditional magnetic inductive loop detectors to radar Doppler systems, and now to computer vision speed measurement, demonstrating a trend towards diversification [2].

The application of magnetic induction loops progressed following the discovery of electromagnetic induction. British physicist Michael Faraday discovered this phenomenon in 1831. By the mid-20th century, these loops were employed in rail transport to monitor the speed and position of trains. As automobiles became widespread and traffic management advanced, their use expanded to other areas of traffic speed measurement. Recent improvements in loop detector technology have made these systems one of the most extensively used vehicle detection technologies due to their low cost, high reliability, and precision [3–5]. However, because inductive loops are embedded in the roadway, they pose challenges in maintenance and are vulnerable to damage.

In 1935, British scientist Robert Watson-Watt developed the first radar, which significantly advanced during World War II. Post-war, radar technology transitioned to civilian applications, including traffic management and law enforcement for vehicle speed measurement. By the 1960s, developments in the aerospace industry greatly enhanced radar precision, extending its operational wavelengths from short waves to millimeter waves, infrared, and ultraviolet. By the 1980s, theoretical research in radar technology had advanced, forming significant theories like radar matched filtering, statistical detection, and ambiguity functions [6]. Today, traffic radar commonly uses the Doppler frequency shift method to calculate vehicle speeds. This technology, which operates wirelessly, minimizes equipment wear and damage and is unaffected by weather conditions, allowing it to function in the dark or under adverse weather conditions. However, in environments with dense traffic across multiple lanes, distinguishing and tracking multiple targets can become complex. Additionally, interference from other electromagnetic devices can increase radar noise and complicate the accurate identification of targets.

In 1960, Theodore Maiman at Hughes Laboratories developed the first operational laser, which laid the foundation for the concept of Light Detection and Ranging (LiDAR). By the late 1970s, NASA had successfully produced an airborne oceanographic LiDAR system equipped with scanning and high-speed data recording capabilities. In 1989, Jeremy Dunn of Laser Technology Inc. introduced a police LiDAR system for traffic speed measurement and law enforcement, significantly improving enforcement efficiency at that time [7]. This system operates by emitting a laser pulse towards a target, capturing the reflected laser, and measuring the time interval. The speed of the target is then calculated by analyzing the time difference between consecutive laser pulses in relation to the speed of light. LiDAR is also critical for acquiring depth information, playing a significant role in 3D scene reconstruction and high-precision mapping [8]. However, LiDAR's performance can be severely affected by adverse weather conditions such as rain, fog, and snow, which scatter the laser beams, and the high production costs of LiDAR limit its large-scale deployment.

The GPS system was initiated by the U.S. Department of Defense in 1970, with its first satellite launched in February 1978, and became fully operational in 1994. Apart from the U.S. GPS, other countries have subsequently developed their own satellite navigation systems, including Russia's GLONASS, Europe's Galileo, and China's Beidou. In 1983, the U.S. decided to declassify the civilian signal of the original GPS system, thereby offering global civilian GPS navigation services and promoting the development and application of civilian GPS receivers. By the 1990s, nearly every car manufacturer had begun to explore and test GPS for navigation and speed measurement capabilities. In 1990, Mazda released the first vehicle with an integrated GPS system, and by the 21st century, vehicle navigation and speed measurement based on the Global Positioning System had become a standard feature in vehicles. The Global Positioning System calculates the receiver's position using signals from multiple satellites and the time differences between them, and the receiver's average speed is then determined from the positional changes over time intervals [9]. GPS is known for its excellent speed measurement accuracy but is susceptible to signal interference from buildings and is only suitable for measuring the speed of the device itself.

In 1959, neurophysiologists David Hubel and Torsten Wiesel conducted vision experiments on cats, discovering that neurons in the primary visual cortex are sensitive to moving edges. This discovery, revealing the columnar structure of visual processing in the brain, would later influence the development of convolutional neural networks four decades later. In 1965, Lawrence Roberts published "Machine Perception of Three-Dimensional Solids," describing the process of deriving three-dimensional information from two-dimensional images. This work fostered the development of edge detection algorithms and three-dimensional reconstruction techniques, providing foundational insights for the field of computer vision.In the realm of speed measurement, two prevalent methods for detecting moving objects in video

sequences are frame differencing and optical flow. Frame differencing was first proposed by Lucas-Kanade in 1981 [10], and optical flow was introduced by Horn-Schunck in the same year [11]. These methods start by processing consecutive frames through frame differencing, optical flow, or background subtraction [12] to extract the motion pixels of the objects. The 3D coordinates of the scene's targets are then calculated using pre-calibrated camera parameters, and the object's speed is determined by analyzing the time differences.Since Alex Krizhevsky, Geoff Hinton, and Ilya Sutskever won the ImageNet competition in 2012, deep learning has brought significant breakthroughs to computer vision. The advent of R-CNN in 2014 pushed object detection algorithms towards an end-to-end approach[13], reducing manual intervention in image recognition. As the fields of computer vision and deep learning have evolved, the measurement accuracy of visual speed measurement now complies with the national standards for electronic speedometers (<=±5% or ±5 km/h), and extensive domestic and international highway camera speed measurement experiments have confirmed the feasibility of computer vision-based speed measurement [14]. Unlike systems based on inductive loops or radar, computer vision speed measurement technology primarily relies on standard cameras and computing platforms, which can extract rich information from images and are relatively easy to deploy and expand, facilitating software upgrades and functionality enhancements. However, variations in lighting, weather conditions, and transitions from day to night can affect image quality, subsequently impacting the accuracy of speed measurement.

This investigation focuses on comparing various sensors' speed measurement methods and technologies, including fixed and vehicle-mounted speed measurements. It analyzes the research directions of speed measurement over different periods in chronological order, categorizes the speed measurement methods and technologies used, and assesses the characteristics and advantages of vehicle speed measurement in the visual field. The paper also evaluates the performance of current technical literature results.

The structure of the paper is as follows: Section 2 provides an overview of common vehicle speed measurement methods and introduces the principles behind various speed measurement techniques. Section 3 introduces existing measurement technology indicators and evaluation methods, comparing the performance of different speed measurement technologies. Finally, Section 4 concludes the paper and looks forward to future research directions in the speed measurement field.

2. Speed measurement technology principles

2.1. Inductive loop

Inductive loops are currently the most mature method for vehicle speed measurement. They employ two methods to detect vehicle speeds. One method involves embedding the loops either underground or directly beneath the object being measured, utilizing changes in the signals from two induction coils triggered by the object to calculate the vehicle's speed [3, 5, 15, 16]. Specifically, when a vehicle passes over an inductive loop, the metal chassis of the vehicle intersects the magnetic lines of force. According to the principle of electromagnetic induction, this interaction induces a signal at the ends of the coil that exhibits regular changes in amplitude and phase. This induced signal is then processed to determine the vehicle's relative position and speed. The principle behind this technique is illustrated in Figure 1.

The other method involves using sensors installed on the axle, which calculate speed based on changes in electrical signals caused by the Hall effect as the vehicle's wheels drive a speed-measuring gear [17]. This approach offers the advantage of obtaining speed data in real-time and represents a new direction in research within the field of inductive loop speed measurement. It is commonly used for measuring the speed of trains. The process of this method is illustrated in Figure 2.

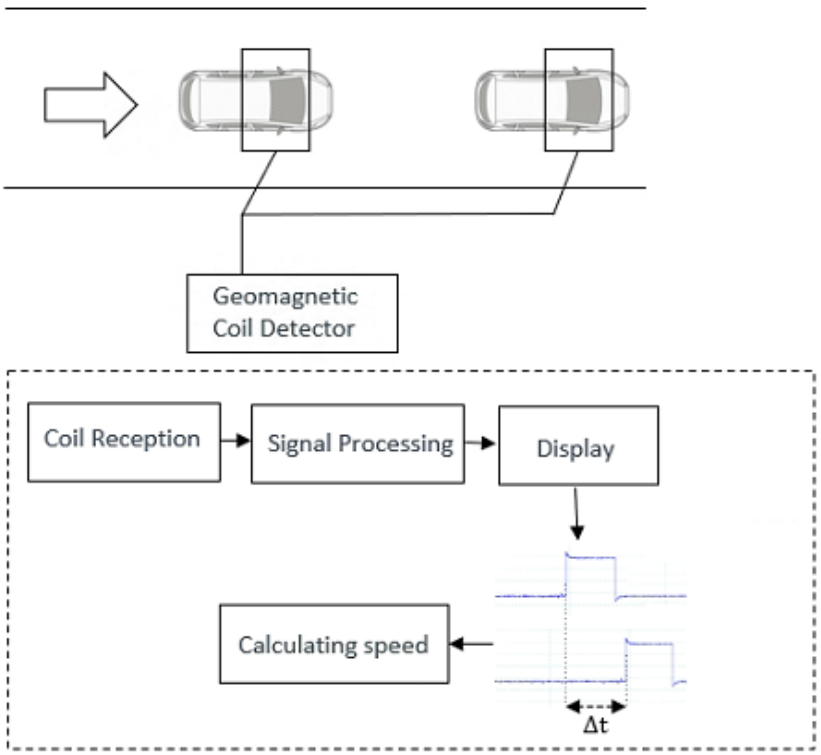

Figure 1. Process diagram of vehicle speed measurement using a dual inductive loop system.

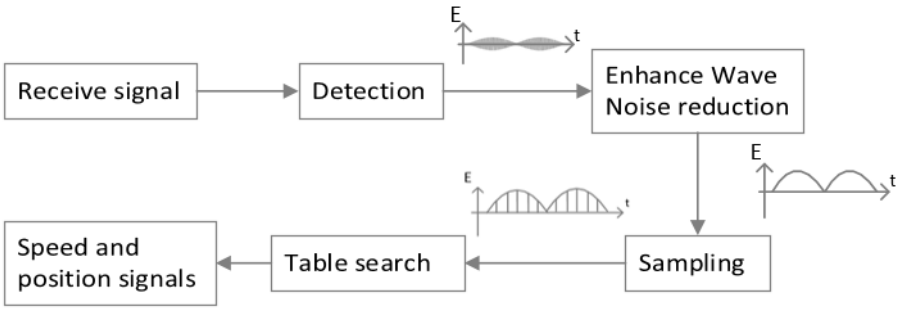

Figure 2. Flowchart of the induction signal demodulation and lookup table approach.

Due to the maturity of inductive loop technology and its accuracy in speed measurement, it is often used as a reference standard for vehicle speed in research fields to assess the accuracy of other speed measurement technologies [18].

2.2. Wireless radar

Wireless radar, particularly millimeter-wave radar, is commonly used for motor vehicle speed measurement. The application of radar in speed measurement often employs the Doppler principle [19]. The Doppler frequency shift can be represented by the following formula:

$$f_d = \pm \frac{2v\cos\theta}{\lambda} \tag{1}$$

f_d is the Doppler frequency shift, v is the relative velocity between the wave source and the receiver, θ is the angle of deviation in their relative position, and λ is the wavelength. The

sign is positive when the wave source is approaching the receiver, and negative when it is moving away.

The error in Doppler radar speed measurement generally does not exceed 1% [6]. The typical installations for radar speed measurement are as follows: Roadside inclined installation is commonly used for experimental or temporary road speed measurement setups[6, 20–24]. This method is convenient for installation and removal but can be prone to obstruction and interference in high traffic volumes. Another method involves mounting the radar on overhead gantries or signal frames [25–28], which reduces the impact of vehicle obstruction but is more challenging to install. Additionally, it requires frequency modulation to separate the speeds of different vehicles [26]. The final method is vehicle-mounted installation [17, 29–31], typically used for measuring the vehicle's own speed from the bottom. However, this approach can be significantly affected by environmental interference [29]. A dual radar design at the vehicle's bottom can enhance measurement accuracy, with differential common error control keeping the error within 1% [30, 31]. There are also front-mounted radars for measuring the relative speed of vehicles ahead [19, 32]. In addition to Doppler frequency shift-based speed measurement, indirect speed measurement using the echo principle is also employed.

$$R = \frac{\mathrm{ct}_r}{2} \tag{2}$$

R is the distance between the radar and the target, c is the speed of light, and t_r is the time difference between the emission of the radar signal and the receipt of the echo signal. Speed is measured based on the relative position change of at least two echoes [33, 34]. The deviation can be greater than 10%, and it is generally not used for speed measurement. Instead, it is commonly applied in autonomous driving for vehicle or object detection [35, 36].

2.3. LiDAR

Similar to radar, LiDAR (Light Detection and Ranging) speed measurement also uses Doppler frequency shift and reflection time methods. Given LiDAR's capability to effectively reflect the contours of objects and assess distances, it is primarily used in the field of autonomous driving. Mounted around the vehicle, it identifies objects in the vicinity and can be used for 3D modeling of the surrounding environment. A common speed measurement method uses the centroid of the point cloud clusters as the vehicle's position. As a vehicle passes the LiDAR sensor, the centroid's relative position changes frame by frame. By analyzing these changes, the speed of other vehicles can be estimated with an error within 2 km/h [37, 38]. There are also methods that modulate the LiDAR to achieve Doppler frequency shift speed measurement, with an error margin around 4% [39–41]. Commonly used LiDAR speed guns emit multiple laser signals and estimate the target's speed based on the reflection time, with a speed error within 2 km/h [42].

Multi-object recognition from point clouds is a crucial step in vehicle speed measurement. With the advancement of artificial intelligence in autonomous driving, approaches for object detection using point cloud data are categorized into three types. The first category [43–45] follows the methodology of PointNet [46] by directly extracting features from raw point cloud data. The second approach converts point cloud data into 2D Bird's Eye View (BEV) images. Although this top-down projection loses height information and may distort object shapes, it reduces computational load [47–49]. The third category involves replacing sections of the point cloud with voxels and then using 3D convolution to extract features, although this transformation can lead to some loss of information [50–54]. As shown in Figure 3.

The first category involves point cloud convolution prediction, focusing on how to sample high-quality target point cloud information. PointNet++ [55] addresses issues related to point cloud density by optimizing feature extraction using the Farthest Point Sampling (FPS) algorithm and Multi-Scale Grouping (MSG), simultaneously employing a feature pyramid

structure to merge shallow and deep features. 3DSSD[56] also improves the sampling method of point clouds. Its Fusion-FPS (F-FPS) filters out background points while retaining more foreground points, removing the Feature Propagation layer to accelerate model computation.

The second category is 2D BEV prediction, where a major challenge is how to retain 3D spatial information in 2D effectively. In the PointPillars network [57], the Pillar Feature Net concept preserves the average height information of point clouds while projecting them onto a plane, followed by object detection using CNN+SSD, achieving a good balance between speed and performance. RT3D [58] encodes three-dimensional point cloud information and projects it into a BEV image for object detection, managing to process point cloud scan data in real-time.

The third category is 3D voxel prediction, focusing on two research objectives: reducing the computational overhead of 3D convolution to speed up processing and obtaining better voxel information to improve accuracy. VoxelNet [59] uses voxels to segment the point cloud in space, randomly sampling data within each voxel to save on computational costs, and uses 3D convolution to extract feature information for regression and classification, providing an effective framework for 3D point cloud recognition. Recognizing that point clouds are often distributed on a single directional surface of objects, the Fully Sparse TRansformer (FSTR) introduces a Gaussian weighting algorithm that optimizes traditional sparse voxel center point sampling methods[60]. It enhances model performance by better predicting target center points and refining denoising queries. This method shifts the uniform distribution of noise to a Gaussian distribution, more closely simulating real data distributions. The model is trained to ignore this added noise, thus predicting target bounding boxes more accurately.

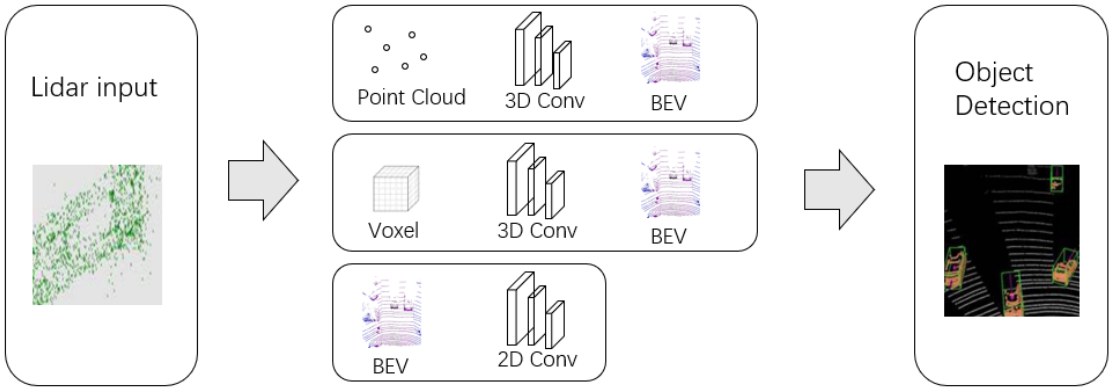

Figure 3. LiDAR object detection process.

In addition to model categorization, there are several universal methods for enhancing accuracy. One such method is proposed by CenterPoint [61], which introduces an anchor-free center point prediction approach. This method locates the center of an object and uses the central features to regress to a complete 3D bounding box, improving accuracy in both 2D BEV prediction and 3D voxel prediction. Given that the distribution of background data in point clouds far exceeds that of the foreground, and that a vast amount of irrelevant information can be detrimental to model performance, Real-Aug [62] approaches this issue from a data scale perspective. It proposes an effective data augmentation technique that generates synthetic foreground point cloud data while ensuring realistic scene logic, thereby enhancing the model's ability to predict foreground targets more accurately. FocalFormer3D [63] uses a Muti-stage Heatmap to identify false-negative samples from a previous phase as challenging samples for focused training in the subsequent phase, while ignoring true positives. This approach helps reduce the model's interference from extensive background information, allowing it to concentrate more on locating foreground objects.

2.4. Global Navigation Satellite System (GNSS)

GNSS speed measurement is typically used for measuring a vehicle's own speed. There are four main GNSS speed measurement methods: position differencing, raw Doppler observation, pseudorange differencing, and carrier phase differencing. Position differencing can achieve high positioning accuracy, usually at the sub-meter level or even higher. The actual accuracy depends on various factors, including the distance between the base station and the rover, the accuracy of the base station, the performance of the receiver, and environmental conditions. Typically, pseudorange relies on the relative differences between receivers at different locations receiving signals from at least four of the same satellites, calculating the spatiotemporal information (x, y, z, t) of the target's position, and using these differences to estimate the target's motion speed. This method usually achieves sub-meter accuracy and is suitable for low-cost receivers [64]. Position differencing technology can achieve several times to several tens of times the accuracy of single receiver positioning, providing more precise and reliable positioning results [9]. Compared to other methods, the original Doppler shift can directly estimate the speed. The mathematical model for the original Doppler shift can be expressed as:

$$D = e\left(v - v_j\right) + b' + \varepsilon \tag{3}$$

Compared to other methods, raw Doppler frequency shift can directly estimate speed. The mathematical model of raw Doppler can be represented as follows: D is the Doppler frequency shift observation value, e represents the direction cosines of the line connecting the receiver and the satellite $\begin{bmatrix} e^x & e^y & e^z \end{bmatrix}$, v and v_j are the velocities of the receiver and the satellite, respectively, b' represents the drifts of the receiver clock and the satellite clock, and ε is the observation noise caused by various error factors such as satellite clock errors, ionospheric errors, and tropospheric errors. In open terrain without interference, raw Doppler observation speed measurement can achieve centimeter-level accuracy [65]. However, in practice, accuracy might be affected by environmental interference and the precision of the receiver. Generally, raw Doppler observation speed measurement can provide relatively accurate speed measurements within a sub-meter range [66]. Carrier phase differential speed measurement involves receiving carrier signals from satellites and using the corresponding receiver module to demodulate the signals, extracting the phase information of the carrier signals and comparing it with locally generated reference signals to measure the carrier signal phase information. The processing of this information can achieve millimeter-level accuracy [64]. Signal precision is enhanced from noise through Kalman filtering [67]:

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k \tag{4}$$

$$P_{k|k-1} = F_k P_{k-1|k-1} {F_k}^T + Q_k \tag{5}$$

$$K_k = P_{k|k-1} {H_k}^T \left(H_k P_{k|k-1} {H_k}^T + R_k\right)^{-1} \tag{6}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \left(z_k - H_k \hat{x}_{k|k-1}\right) \tag{7}$$

$$P_{k|k} = \left(I - K_k H_k\right) P_{k|k-1} \tag{8}$$

The formulas (4) and (5) represent the prediction equations: where formula (4) is the predictive state estimation equation $\hat{x}_{k|k-1}$ predicting the state at time k without considering the measurement z_k , F_k is the state transition matrix, B_k is the control input matrix, and u_k is the control input. Formula (5) is the predictive error covariance, $P_{k|k-1}$ is the corresponding predictive error covariance, and Q_k is the process noise covariance matrix. Formulas (6), (7), and (8) represent the update equations: Formula (6) is the Kalman gain, K_k is the Kalman gain, H_k is the observation matrix, and R_k is the observation noise covariance matrix. Formula (7) updates the state estimation, and formula (8) updates the error covariance. Using differential techniques with these four measurement technologies can eliminate common errors such as

satellite clock errors, ionospheric delay errors, and tropospheric delay errors to a certain extent, stabilizing speed measurement accuracy at the centimeter or millimeter level [68]. Professional GNSS speed measurement devices can serve as a standard reference for speed.

2.5. *Computer vision*

Computer vision for speed measurement typically involves two stages, As shown in Figure 4.

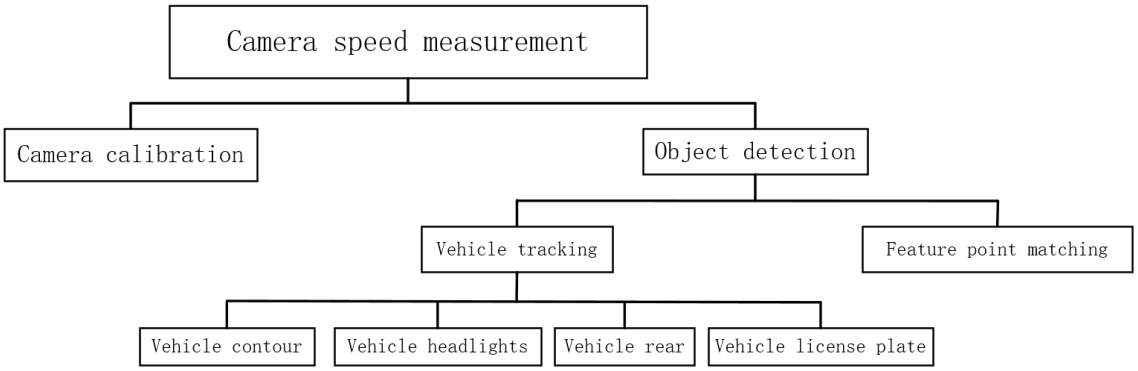

Figure 4. Computer visual vehicle speed measurement flowchart.

Accurate camera calibration is crucial for speed measurement, as the transition from a two-dimensional pixel coordinate system to a three-dimensional world coordinate system determines the precision of speed measurement. The transformation formula for a pinhole camera from a two-dimensional plane to three-dimensional coordinates is:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} (R|t) \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{9}$$

(u,v) are the pixel image coordinates of the target point, f_x and f_y are the camera's focal lengths,c_x,c_y are the optical centers of the image, and$(R|t)$ is the external parameter matrix, including the rotation matrix R and the translation matrix t, used to transform points from the world coordinate system to the camera coordinate system, (X,Y,Z) are the coordinates of the target point in the world coordinate system.

Because the external parameter matrix varies with the scene, calibration for speed measurement in a monocular setup generally falls into two categories: direct scene calibration and reference object calibration. Scene calibration [69] refers to direct calibration in the specific scene. The other method involves objects within the line of sight that have standard or known lengths, from which feature points are extracted to calculate the camera's external parameters. Since depth information cannot be directly obtained in a monocular setup, it is necessary to use the orthogonal relationships within the scene and the vanishing points of lines to establish a three-dimensional coordinate system.First, it's essential to detect the "rectangular areas" contained in the image. By using the parallel relationships of the edges of these "rectangular areas," the vanishing points of the lines are determined. Calibration is then performed based on the orthogonal relationships of the adjacent edges of the "rectangular areas" and the properties of the vanishing points. Since the projection of actual rectangular areas in the scene onto the image plane usually results in irregular quadrilaterals, the key to detection lies in identifying the corresponding areas in the image based on the known orthogonal and parallel relationships in the actual scene. The corresponding area in the image is the "rectangular area." In vehicle speed measurement, the most commonly used reference is the lane's dashed lines, which

usually adhere to uniform standards [70, 71]. Another good option for establishing a coordinate system is using license plates [72]. There are also techniques for real-time automatic calibration of the scene, but these involve significant computational effort [73].

$$v = \frac{\Delta S}{\Delta T} \tag{10}$$

Speed measurement typically uses virtual loops as triggers. The principle of virtual loop speed measurement is divided into two types: measuring the time difference as vehicles pass through virtual loops twice to estimate speed [74, 75], and schemes using virtual frames for multi-lane detection [70, 76], as well as methods using multiple virtual loops [74] to enhance robustness. In Dahl M's experiment, it was proven beneficial to increase the accuracy of speed measurement by increasing the number of intrusion lines; using the same method, increasing from two to four lines reduced the error from 1.92% to 1.17% [77]. Speed is determined by timing the changes in the vehicle's position relative to the virtual loop over a fixed time [78].

In contrast to the pseudo-three-dimensional coordinates obtained with a single camera, binocular recognition can derive true three-dimensional coordinates through the triangulation of target points using two cameras.

$$p = MP = K(I\ 0)P \tag{11}$$

$$p' = M'P = K'(R\ t)P \tag{12}$$

The target point's pixel coordinates in the first camera are denoted as p, with M representing the projection matrix, and K as the camera's internal parameters. P is the three-dimensional coordinate, with p' , M' , K'corresponding to the second camera's parameters. The task then transforms into solving the fundamental matrix F for pixel correspondence. The poles e and e' of the two images are aligned on the same line, and the transformation matrices H and H' are minimized as shown in Equation (13).

$$\sum_i d\left(Hp_i, H'^{p_i'}\right) \tag{13}$$

Image resampling corrects the images to parallel views, as illustrated in Figure 5.

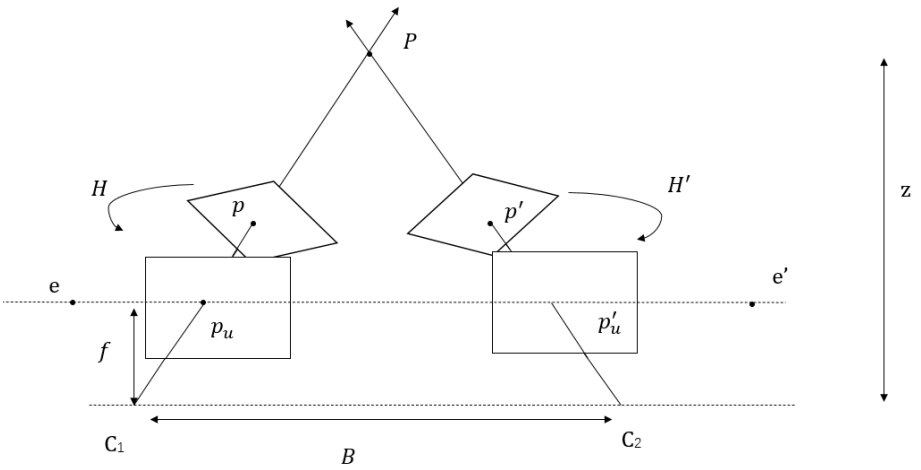

Figure 5. Binocular vision depth estimation.

Finally, the depth information of the target is confirmed by calculating the disparity between the two views, as represented in Equation 14.

$$p_u - p_u' = \frac{B \cdot f}{z} \tag{14}$$

To enhance the robustness of parameters or target points, a series of fitting algorithms such as RANSAC, Levenberg-Marquardt method to minimize reprojection error, or least squares method are often used. These algorithms fit multiple points' data to eliminate erroneous points and obtain highly robust data.

There are two main types of methods commonly used in target recognition tasks: traditional image processing algorithms and artificial intelligence image processing. In traditional processing algorithms, the frame differencing method calculates the difference between corresponding pixel points in two adjacent frames of a video sequence. If the difference exceeds a set threshold, it is considered that motion is present, and the corresponding pixel points are extracted as moving targets [79, 80]. In the field of traffic, where cameras are often fixed, improved algorithms such as background subtraction are commonly used. This method first obtains a background model, then calculates the difference between the current frame and the background model, and extracts different pixel points as moving targets. Popular algorithms include Gaussian Mixture Model for background modeling and Vibe background modeling, among other improved Gaussian background modeling methods. However, background modeling algorithms have obvious limitations, such as the inability to eliminate shadows from objects. They often need to be used in conjunction with shadow removal algorithms, like those based on HSV or RGB color spaces, to remove shadows and obtain accurate motion area images [81]. Another method is optical flow, which identifies objects with similar optical flows by comparing the optical flow of consecutive frames. Common methods include Lucas-Kanade (LK) optical flow and pyramid optical flow. Although optical flow can accurately capture information about moving objects, its disadvantage is that it requires a constant background light and can only detect small-range moving objects, with effectiveness decreasing as the calculation range increases.

When AlexNet achieved first place in the object classification domain of the ImageNet LSVRC-2012 competition with a Top5-error metric exceeding the second-place contender by more than 10% [82], it sparked widespread adoption of deep learning in computer vision, as shown in Figure 6. The proposed techniques in the paper, including ReLU, Local Response Normalization, Dropout, data augmentation, and GPU-accelerated neural network training, have become foundational in modern computer vision. VGGNet [83] introduced the concept of receptive fields, demonstrating that multiple 3×3 convolutional kernels could achieve the effect of a larger kernel with fewer computations. GoogLeNet [84] employed a multi-branch structure to provide the network with more receptive field choices. The use of auxiliary classifiers during training enhanced gradient signal propagation in backpropagation, facilitating better training information dissemination. ResNet [85], a milestone deep learning network in computer vision, introduced a residual structure to address the problem of network degradation in deep networks. This innovation enabled deep learning networks to achieve human-level classification capability [86], and it marked the first implementation of a deep learning network with over 1000 stacked layers.

After R-CNN demonstrated the feasibility of convolutional neural networks (CNNs) for object detection, deep learning-based object detection started to emerge as a new research direction. Within this framework, image classification networks act as the backbone for tasks such as object detection, semantic segmentation, and instance segmentation, facilitating the extraction of image features. R-CNN uses the Selective Search algorithm to generate a series of candidate boxes, then employs AlexNet as the backbone network to classify these boxes and identify objects within the images. In Faster R-CNN [87], a Region Proposal Network (RPN) is utilized, leveraging shared convolutional features to rapidly generate candidate regions, thus avoiding the time-consuming process associated with the Selective Search algorithm and significantly improving detection speed.Subsequently, YOLO enhanced image detection speed using a grid scanning method to calculate confidence levels, albeit with a compromise in accuracy. SSD adapted YOLO's approach of transforming detection into regression and

also incorporated the anchor mechanism from Faster R-CNN. However, unlike Faster R-CNN, where anchors are precisely adjusted at each position, SSD, similar to YOLO, creates anchors on a grid, ensuring both speed and accuracy. DETR [88] utilizes the features of the Transformer architecture, employing an encoder to extract feature information and a decoder to directly output the categories and positions of 100 targets. These 100 items (default values) are then refined through Hungarian bipartite matching in the decoder, optimizing the prediction boxes layer by layer. Unmatched predictions are labeled as 'no object'.

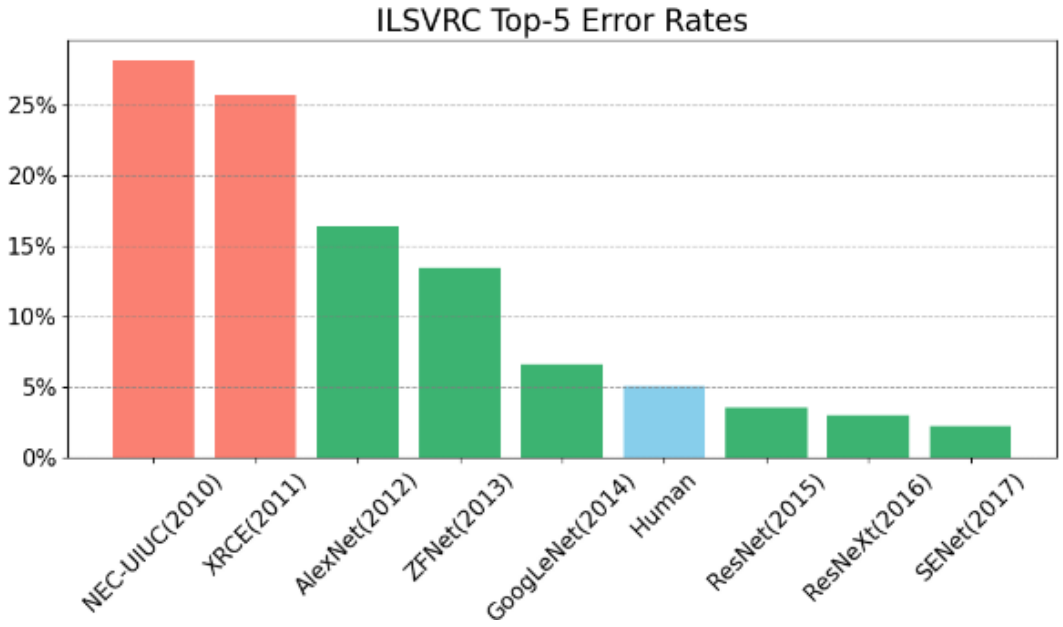

Figure 6. Best model of the year in the ImageNet large scale visual recognition challenge.

Table 1. Comparison of Deep Learning Object Detection Algorithm Performance.

Algorithm/ Standard	YOLOv8 (You Only Look Once)	Cascade Mask R-CNN (Region-based Convolutional Neural Networks)	SSD (Single Shot MultiBox Detector)	Transformer
Processing speed	Fast	Moderate	Faster	Moderate
Bounding box accuracy	High	High	Moderate	High
Parameters	Moderate	High	Moderate	High
Recognition accuracy	High	High	Moderate	High

As shown in Table 1,YOLO and SSD generally outperform Cascade Mask R-CNN and Transformer models in terms of computational speed. However, Cascade Mask R-CNN and Transformer models excel in the precision of bounding boxes and the detection of small objects. Most models can achieve relatively high accuracy in recognizing target types. With iterative updates, deep learning algorithms have improved in bounding box precision and the ability to detect small targets, but the actual performance of these models still correlates directly with their total parameter count.

Transformers, initially proposed by Google for natural language processing tasks [89], are now applied in the field of image recognition. Due to their powerful self-attention mechanism and parallel computing nature, Transformer architectures have been widely used in various fields. The self-attention mechanism of Transformers can perform functions similar to CNNs under specific constraints [90]. Vit-transformer [91] and Swin-transformer [92]

are examples of Transformers specialized for image domains. With datasets larger than 100M, Transformer models exhibit stronger learning characteristics than CNN networks [91]. Moreover, since Transformers were originally designed for natural language processing, they also have inferential capabilities in image processing. They can perform image processing functions such as de-raining [93–95], de-fogging [94–96], image restoration [95, 97], and pixel enhancement [95, 97, 98]. The processed images are clearer for object recognition tasks. Additionally, Transformers have shown effective results in semantic recognition of videos [99] and target ID matching [100]. making them a versatile tool in advanced image processing and analysis.

In recent years, the Transformer model DETR [88] for image detection has evolved rapidly. Deformable DETR [101], inspired by deformable convolution [102], introduces a deformable attention module in the Transformer. This module utilizes a deformable attention mechanism to gather more sampling points, enabling the model to focus on key object features during training and improve convergence efficiency. This structure has enhanced the convergence of DETR from 500 epochs to just 50 epochs. Conditional DETR [103] posits that the slow convergence of DETR is due to the queries needing to simultaneously learn content and spatial aspects, making it challenging for the model to converge. To address this, the authors propose a unique spatial embedding using a concatenation rather than addition approach, allowing the model to focus on different tasks to expedite convergence, improving model efficiency to 108 epochs.DAB-DETR [104] builds on Conditional DETR by further refining the queries with four-dimensional coordinate initialization, optimizing them at every layer of the decoder. DN-DETR [105] follows the four-dimensional coordinate approach of DAB-DETR, addressing inconsistent matching across decoder layers by training each decoder detection directly with Ground Truth (GT). It also incorporates Denoising training to minimize inconsistencies caused by Hungarian matching between different decoders. DINO [106] inherits ideas from DAB-DETR [104] and DN-DETR [105]. Firstly, it modifies the task of predicting real boxes at every layer in DAB-DETR to use cross-attention as a shortcut for learning relative offsets. Secondly, considering that most decoder predictions in practical environments are negative samples, it introduces high-noise negative sample prediction tasks on top of DN-DETR, enhancing the model's ability to differentiate between positive and negative samples.

Co-DETR [107], the current state-of-the-art (SOTA) model in image detection, contends that the original DETR suffers from inefficient query matching and an insufficient number of positive samples, which hampers model training. Similar to GoogLeNet [84], it utilizes multiple auxiliary heads to accelerate training. However, it differs by using varied detection heads trained together, including Faster-RCNN [87], ATSS [108], and RetinaNet [109], further enhancing the accuracy of the SOTA model DINO-Deformable-DETR [106].

Due to the requirements for model size and real-time performance, there is limited literature on using Transformers for vehicle speed measurement. However, Zhao Y *et al.* proposed the RT-DETR [110] network, a hybrid model combining CNNs with Transformers, addressing the slow detection speed of Transformer models. Thus, using Transformers for vehicle speed measurement is a promising research direction.

In object detection tasks, a well pre-trained model can more rapidly assimilate new task features. By transferring models trained on datasets such as COCO to vehicle recognition tasks, models can more quickly master vehicle feature information [111]. Therefore, the following datasets might be used to train a deep learning network for vehicle recognition, as shown in Table 2.

Table 2. Overview of the vehicle detection dataset.

Dataset	Application domain	Resolution sampling rate	Number	Information
Microsoft COCO [112]	Image	640*480	330,000 images (37.57 GiB)	80 Labels Bounding box
BrnoCompSpeed [113]	Speed camera	1920*1080 50Hz	18 h (180GiB)	Three angle of viewsCars' Video-time and Speed
UTFPR-HSD [114]	Traffic video/image	1920*1080 25Hz	15664 frames (10GiB)	6 Labels Frame and numbers
QMUL junction [115]	Traffic video	360*288 25Hz	1h(324MiB)	None
Vehicle Speed Measurement (UTFPR) [116]	Speed camera	1920*1080 30Hz	20h(30GiB)	License plate's B-box Car duration frames speed
KITTI [117]	Traffic images	1240*370	30,000 images (180GiB)	Calib 9 Labels bounding box
kinetics-700 [118]	Videos	452*256 30Hz	650,000 videos (700GiB)	700 Labels Duration times
The FLIR Thermal Starter Dataset [119]	Thermal traffic video/images	640*512 24Hz	26,442 frames (3.5GiB)	15 Labels Bounding box
HIT-UAV [120]	UAV Thermal images	640*512	2898 frames (814MiB)	5 Labels Bounding box Altitude and Camera perspective
nuScenes [121]	Traffic images	1600*900	1.4 Million (547.98GiB)	23 Labels Bounding box
Waymo [122]	Traffic images	1920*1280	12 Milion (1TiB)	23 Labels Bounding box

In addition to model self-improvement through transfer learning from larger training sets, Hinton G *et al.* [123] have proposed a method known as knowledge distillation, which involves training a smaller student model using a larger teacher model.This method combines the results of the large model with actual outcomes to compute a weighted sum loss, effectively aiding the training of the smaller model, as illustrated in Figure 7. Knowledge distillation can enhance the performance of smaller models, reduce the resource consumption for model deployment, and potentially accelerate inference speeds. For instance, eva-2 [124] utilizes the previous generation multi-modal eva model [125] as a teacher model and, through knowledge distillation [126], achieves better results on multiple vision task datasets while reducing parameter size.

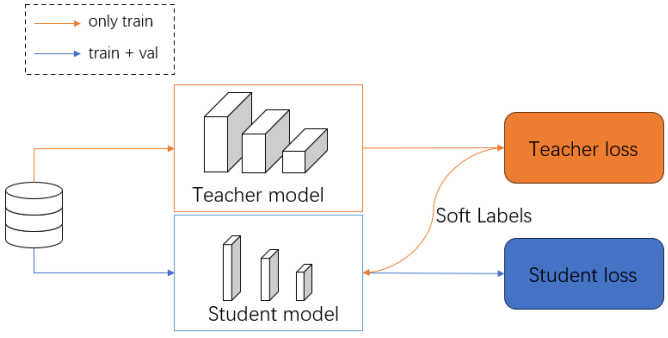

Figure 7. Knowledge distillation process.

Supervised learning requires large amounts of calibrated data from various datasets, whereas self-supervised learning does not require extensive specific task annotations. This approach can enhance model performance and generalization capabilities in scenarios where data annotation costs are high or annotated data is scarce. Self-supervised methods also significantly increase the amount of data that can be trained. After achieving good results in supervised learning, large models are trained using self-supervised methods to leverage massive amounts of data to learn the intrinsic structure and features of the data. For example, GPT-3 [127] is a large model that has demonstrated strong generalization capabilities across many tasks. Using large models like GPT or BERT can easily adapt to downstream tasks through few-shot or zero-shot learning methods. These models, when effectively integrated with other modules, can also be employed for image generation tasks, such as DALL-E [128].

Large vision models are an important trend in recent years, with the multi-modal large vision model CLIP [129] consisting of an image encoder and a text encoder, trained on a vast number of image-text pairs to understand image content and generalize to classify and understand unseen images. The paper highlights that CLIP exhibits a stronger generalization capability than ResNet101 trained on ImageNet, showing over 35% higher accuracy in image classification across datasets such as ImageNet-R, ObjectNet, ImageNet-Sketch, and ImageNet-A. Similarly, SAM [130] is a large vision model for image segmentation that uses "prompt" technology for zero-shot or few-shot learning on new datasets and tasks. SAM has proven extremely effective in zero-shot transfer learning, outperforming the previous RITM [131] on most of the 23 datasets evaluated. By modifying the model structure to freeze the backbone network, large models can effectively transition to downstream tasks. For instance, MedSAM employs an Adapter mechanism to transfer SAM to the medical field, achieving state-of-the-art (SOTA) results in most medical tasks [132]. In the field of autonomous driving, A recent line of studies [133, 134] utilizes SAM to segment foreground and background scenes on roads to aid in point cloud computation. RegionSpot [135] combines SAM and CLIP by using cross-attention between objects segmented by SAM and the image features of CLIP to regress on segmentation categories.

For the objects identified, the subsequent task is object detection. The first method involves detecting the identified objects using the centroids of their contours [70, 136]. While this can reduce the computational load and improve operational speed during detection, centroids are more susceptible to occlusion. More commonly, corner detection algorithms are used for object detection, such as the Harris corner detection algorithm and the SIFT corner detection algorithm. Corners typically have distinct features on objects, exhibiting significant grayscale changes compared to surrounding areas, such as on vehicle contours, headlights, or the corners of license plates. Effective detection of target objects can be achieved through matching these corners [78, 116, 137, 138].The second method [139–143] involves using the Hungarian algorithm to match vehicle identification frames on two frames that have undergone recognition algorithms, which can also effectively detect targets. The third method is based on deep learning [144–150], utilizing Convolutional Neural Networks (CNNs) for object detection, and association networks such as Graph Convolutional Networks (GCN) or sequence networks (RNN, LSTM) for detecting objects across different frames. The success of Transformers in sequence and image tasks has demonstrated their capabilities in the field of object detection. Using an encoder for image detection and a decoder for sequence detection shows promising results.

2.6. Multi-sensor fusion

In complex applications such as autonomous driving, data fusion can improve the accuracy of object detection and the reliability of speed measurement. Multi-sensor fusion techniques ensure performance that is on par with or even surpasses that of the optimal single sensor under most conditions [151]. One of the earliest multi-sensor fusion techniques used in the

transportation sector is the integration of Inertial Measurement Units (IMU) with Global Positioning Systems (GPS), where GPS usually provides precise location data in open spaces but suffers from decreased accuracy in signal-obstructed environments. Meanwhile, IMUs can accumulate errors over time without external correction. Dynamic data fusion through algorithms such as the Kalman Filter and Particle Filter greatly enhances the reliability of positioning [152].

In terms of environmental perception, commonly used sensors include LiDAR, Radar, and cameras. Multi-sensor fusion includes Camera-LiDAR (CL), Camera-Radar (CR), LiDAR-Radar (LR), and Camera-LiDAR-Radar (CLR) fusion approaches. The advancement in the field of artificial intelligence has also facilitated progress in multi-sensor fusion, as shown in Figure 8, with deep learning enabling deep-level feature fusion of LiDAR point clouds, camera images, and Radar point clouds to enhance object detection performance. For example, the single-sensor SOTA model FocalFormer3D [63] achieves an mAP of 70.5%, which is weaker than the CL fusion SOTA model BEVFusion4D [153] with an mAP of 76.8%.

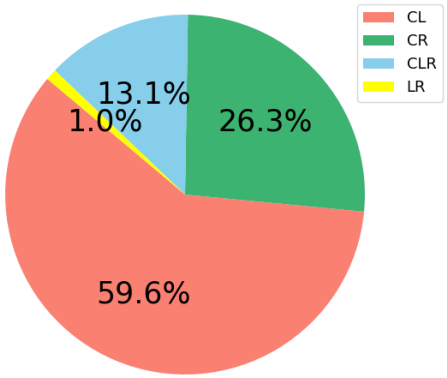

Figure 8. The proportion of autonomous driving research based on deep learning multi-sensor fusion over the past five years.

Camera-LiDAR-Radar Fusion, CLR fusion allows the model to incorporate the characteristics of three different sensors. Ratheesh Ravindran *et al.* [154] proposed a CLR multi-sensor fusion Bayesian Neural Network (BNN) model. This model quantifies the uncertainty of predictions from various sensors by training on the probability distribution of the model's parameters. It also demonstrates the effectiveness of different sensor fusion outcomes through ablation studies and finds that, in most cases, CL fusion performs better than CR fusion, though CR fusion outperforms CL in adverse conditions like rain, snow, and high winds.

Camera-LiDAR Fusion, LiDAR offers better imaging resolution than Radar, allowing for better differentiation of vehicle features in dense traffic, while also compensating for the errors in depth information calculated from camera images. There are two models named BEVFusion [155, 156], each employing different fusion strategies. Liang T *et al.* [155] use a decision-level fusion strategy, processing point cloud and image data separately before merging results with a fusion module trained on three prediction maps, allowing the model to continue functioning if one sensor fails, hence increasing robustness. Liu Z *et al.* [156] extract features from images and LiDAR and merge them on the BEV for faster processing. TransFusion [157] uses a cross-attention mechanism at the feature level, merging key point cloud data guided by images to generate initial prediction maps, then combining critical image information for decision-level fusion to produce final results.

Camera-Radar Fusion, Although Radar is less expensive, less affected by environmental factors, and capable of long-range detection, its image resolution is low and the collected point clouds are sparse. It generally assists in image processing during feature fusion. In

HVDetFusion [158], First, model uses multi-camera fusion to identify candidate boxes for targets, then filters out Radar point cloud data that deviates significantly from these boxes. Lastly, spatial information about speed, direction, and position from Radar is integrated into the candidate boxes to enhance accuracy. CRN [159] employs a multimodal variational attention mechanism to fuse information from both sensors at the feature level, addressing the issue of spatial misalignment during fusion.

LiDAR-Radar Fusion, Both sensors overlap significantly in functionality; their fusion primarily aims to adapt to harsh environments and enhance the robustness of detection algorithms. Bi-LRFusion [160] and CenterPoint-Ensemble [61] fuse features on both BEV and Voxel, showing a 2%–3% improvement in mAP compared to models using only LiDAR.

3. Speed measurement performance

3.1. Development trends

By researching over a hundred pieces of domestic and international literature from different periods, the following chart has been created to illustrate the development and changes in various sensor technologies over time.

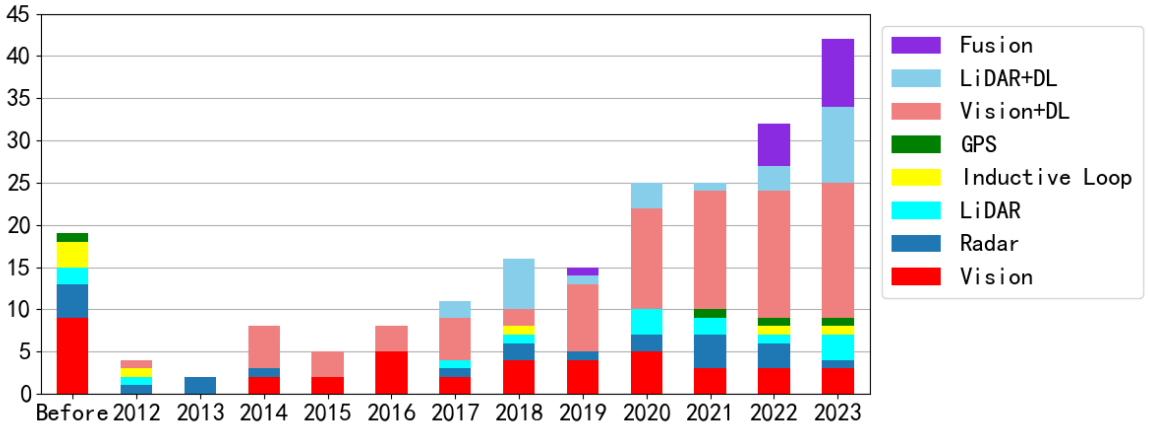

Figure 9. Different vehicle speed measurement technologies research trend chart.

Through Figure 9 classification, it can be observed that research in the field of motor vehicle speed measurement using computer vision has shown a clear upward trend after 2014. During this time, the RCNN model and Faster RCNN's high accuracy in image recognition sparked interest in the study of convolutional neural networks for image recognition [13, 161]. Currently, most researchers have a greater inclination towards exploring the field of visual speed measurement.

3.2. Evaluation metrics

In the field of satellite speed measurement, commonly used evaluation metrics include RMS (Root Mean Square). used to evaluate the deviation of speed prediction.

$$RMS = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{\mathrm{N}}} \tag{15}$$

Satellite positioning accuracy is crucial for subsequent speed measurements. CEP (Circular Error Probable) is frequently used to assess model positioning accuracy in satellite applications.

$$CEP = k\left(\sigma_x + \sigma_y\right) \tag{16}$$

Here, k is a coefficient related to the confidence level and associated with the quantiles of the normal distribution, and σ is the standard deviation of positioning errors. Different k values correspond to different accuracies, with statistical measures ranging from small to large as CEP, CEP95 (R95), and CEP99. These represent a circle with (μ_x, μ_y) as the center and CEP, CEP95, CEP99 as radii, indicating the probability of points falling within the circle as 50%, 95%, and 99%, respectively.

In the field of computer vision image processing, common evaluation metrics for assessing picture recognition accuracy include the following:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{17}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{18}$$

In vehicle testing, TP (True Positive) indicates the number of vehicles correctly identified, FP (False Positives) refers to non-vehicles misclassified as vehicles, and FN (False Negatives) represents vehicles misclassified as non-vehicles. Precision in image domain focuses on accuracy, and Recall on the number of identifications. Many image detection datasets use IoU (Intersection over Union) as an evaluation criterion. IoU is a metric for measuring the overlap between two sets.

$$IoU = \frac{|B_r \cap B_p|}{|B_r \cup B_p|} \tag{19}$$

B_r is the true Bounding box of the object, and B_p is the predicted Bounding box. The IoU value ranges between 0 and 1, indicating the degree of overlap between two sets. An IoU of 1 means complete overlap, and 0 indicates no overlap. In object detection tasks, IoU is typically used to evaluate the overlap between detection boxes and real target boxes to determine the accuracy of detection. Common IoU thresholds are used to assess whether a detection box has correctly identified a target.

$$AP = \int_0^1 \max_{\tilde{r} \geq r} p(\tilde{r})\, dr \tag{20}$$

In object detection tasks, such as those based on the COCO dataset, Intersection over Union (IoU) is commonly used to distinguish recognized objects, thereby calculating Precision and Recall, and ultimately computing the mean Average Precision (mAP). Here, $P(r)$ represents Precision at Recall r. By averaging Precision-Recall curves across N categories, mAP is obtained.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{21}$$

mAP is a commonly used metric in the field of object detection to evaluate model performance. It comprehensively considers both Precision and Recall, and provides a unified measurement of performance across different categories. This composite metric helps understand the model's performance across various categories, thus enabling a more comprehensive evaluation of model performance.

The evaluation metric commonly used to assess the accuracy of speed estimation is the Mean Absolute Error (MAE). Unless otherwise specified, this metric is generally adopted by most speed measurement methods for evaluation.

$$MAE = \frac{1}{n}\sum_{n=1}^{n} |y_i - \hat{y}_i| \tag{22}$$

In terms of speed measurement evaluation standards, internationally, the Maximum Permissible Error (MPE) is commonly used as the criterion for whether speed measuring equipment is

qualified. The formula for the Maximum Permissible Error is as follows:

$$MPE = \begin{cases} \pm x \quad , & S \leq 100km/h \\ \leq x\%, & S > 100km/h \end{cases} \tag{23}$$

S represents the actual speed measurement, and x denotes the maximum allowable error range, which varies by country (e.g., China adopts a value of 6, while many European countries use 3).

3.3. Performance comparison

After researching numerous publications, the following is a compilation of various speed measurement techniques and methods proposed by different scholars, along with the accuracy results obtained using the MAE (Mean Absolute Error) evaluation metric. As shown in Table 3.

Table 3. Comparative study of different vehicle detection and speed measurement techniques.

Author	Vehicles detection	Speed estimation algorithm	Performance (device position speed rangemean error)
Sochor*et al.* [162]	Background Subtraction, Faster-RCNN detect vehicle, Kalman filter tracking	Detect distance within time interval	Gantry, 60–110 km/h, 1.3%
Saleh*et al.* [14]	Optical flow tracking vehicles	Detection time at the same distance (four intrusion lines)	Gantry, 50–130 km/h, 2.17%
Alexander*et al.* [163]	Background Subtraction, Difference frame detection, Mixture of Gaussian tracking	Detect distance within time interval	Gantry, No description, 3.86%
Cheng*et al.* [76]	Background Subtraction, YOLOv4 detect vehicle	Detection time at the same distance (intrusion box)	Gantry, 30–120 km/h, 7.6%
Biswas*et al.* [137]	Faster R-CNN detection, CSRT tracking vehicle	Detection time at the same distance (a section of the way)	UAV, No description, 4%
Ke*et al.* [164]	Kanade-Lucas optical flow tracking vehicle, k-means Distinguish between background and vehicle speed	Detect distance within time interval (five frames)	UAV, 60–90 km/h, 11.6%
Yang*et al.* [138]	Multi-Camera, SSD detect License plate, SURF tracking License plate	Detect distance within time interval	Roadside, 30–60 km/h, 3.8%
Liu *et al.* [165]	YOLOv3 detection, Kalman tracking	Detect distance within time interval(frame)	UAV, 30–80 km/h, 7.1%
Maduro *et al.* [166]	Background Subtraction, Kalman tracking	Detection time at the same distance (two intrusion lines)	Gantry, 50–100 km/h, 2%
Doğan *et al.* [167]	optical flow	Detect distance within time interval(frame)	Roadside, 30–80 km/h, 1.9%
Czajewski*et al.* [168]	SVM-classifer detect license plate, Adaptive thresholding matching OCR tracking	Detection time at the same distance	Gantry, 40–80 km/h, 4%
Li *et al.* [169]	YOLOv3 detect vehicle, Optical flow tracking background, Kalman track vehicle	Detect distance within time interval(frame)	UAV, 20–80 km/h, 1.28%

Table 3. *Cont.*

Author	Vehicles detection	Speed estimation algorithm	Performance (device position speed range mean error)
Luvizon*et al.* [116]	Motion History Image concept detection, KLT and SIFT track vehicle license plate, SVM identify license plate	Detect distance within time interval	Gantry, 10–70 km/h, 3.4%
Zhang *et al.* [139]	Mask R-CNN detect vehicle, SORT and Hungarian algorithm tracking, Wheel build 3D bounding boxes	Detect distance within time interval(frame)	Gantry, No description, 4%
Zhang *et al.* [37]	SVM classification to detect vehicles in point cloud data, Unscented Kalman Filter and Joint Probabilistic, Data Association Filter track vehicle centroid	Detect distance within time interval(frame)	Roadside, 0–50 km/h, 3.2%
Du *et al.* [20]	Radar the principle of interference with multiple receiving antennas	Doppler velocity measurement	Roadside, 50–130 km/h, 0.6%
Nie *et al.* [5]	The loop sensor velocity measurement	Detection time at the same distance	Ground, 40–180 km/h, 3.8%
Ma *et al.* [66]	Dynamic PPP Parameter Method and Carrier Phase-Derived Doppler Velocity Method	Detect distance within time interval	Car, 0–70 km/h, 0.5%

4. Discussion

In this survey, we focused on the various sensor-based speed measurement methods that have been developed in recent years. However, the study revealed that there are still some shortcomings in the vehicle speed measurement technologies used in intelligent transportation systems. Table 4 outlines the advantages and disadvantages of each speed measurement method.

Table 4. Characteristics of different speed measurement methods

Technology	Advantages	Disadvantages
Inductive loop	1) Mature technology 2) Accurate speed measurement	1) Easily damaged by large vehicles. 2) Installation or repair require traffic interruption
Microwave radar	1) Excellent performance in adverse weather conditions, all-weather operation 2) Can detect multiple lanes in a lateral manner 3) Long working distance 4) Accurate speed measurement	1) Large vehicles obstructing adjacent lanes for small vehicles 2) The more lanes, the greater the measurement error 3) Higher installation conditions are required
Lidar	1) Easy and quick installation 2) High range resolution 3) Accurate detect vehicles 4) Accurate speed measurement	1) Only detect a single lane 2) Insufferable for bad weather 3) Costly to implement
GNSS	1) High precision 2) Real-time capability 3) Global coverage	1) The signal is susceptible to environmental interference 2) Not suitable for traffic speed measurement

Table 4. *Cont.*

Technology	Advantages	Disadvantages
Computer vision	1) Provides visual images for accident management 2) No need for road construction disruption 3) Offers a wealth of traffic management information 4) Can detect multiple lanes	1) Large vehicles obstructing adjacent lanes for small vehicles 2) Highly susceptible to environmental influences

Inductive loops, although widely used in traffic flow monitoring, are not sensitive to small or non-metallic vehicles and are prone to malfunction in case of road damage or water accumulation. LiDAR provides precise environmental mapping capabilities, but its high cost and susceptibility to rain and fog affect measurement accuracy. Wireless radar, while efficient in speed and distance detection, is susceptible to interference from building reflections and other electronic devices in urban environments. GNSS systems work well in open areas but suffer from signal blockage and multipath effects in densely built cities or indoor environments, reducing their accuracy and reliability. Computer vision technology, while excellent in vehicle recognition and tracking, is sensitive to lighting changes and requires substantial computational resources for processing high-resolution videos, posing technical challenges in real-time applications.

Through this research, vehicle speed was measured directly or indirectly using data from various sensors, with each technology possessing unique strengths and challenges suited for different application environments. There is an increasing trend towards adopting non-intrusive speed measurement methods, which do not require physical modifications to the infrastructure, Technologies such as computer vision and advanced radar systems are at the forefront of this shift, offering flexible and scalable solutions that can be easily integrated into existing traffic systems. Hybrid methods that blend multiple data sources hold immense potential, multi-sensor fusion combines data from diverse sources such as cameras, radar, and LiDAR, helping to overcome the limitations of individual sensors. Integrating sensors with deep learning networks can further enhance the accuracy and robustness of predictions.

5. Conclusion

This paper introduces various speed measurement methods and technologies, including inductive loops, LiDAR, wireless radar, GNSS, and computer vision, and analyzes their performance, advantages, and disadvantages. An analysis of experimental details and results from over a hundred domestic and international papers on speed detection technology shows that, with the rapid development of artificial intelligence, research combining speed measurement methods like inductive loops, LiDAR, wireless radar, GNSS, and computer vision with artificial intelligence is evolving swiftly.

Therefore, future research needs to focus on optimizing multi-sensor fusion technology, developing low-cost and efficient sensors, and combining deep learning techniques for real-time data processing. Addressing these issues through research will not only improve the accuracy and reliability of speed measurement technology but also drive intelligent transportation systems to higher levels of automation and intelligence. These improvements will provide more accurate tools for traffic management and a safer driving environment for autonomous vehicles.

Acknowledgments

This work was supported in part by the Ministry of Education's Industry-University Cooperative Education Program with grant number 230706093313536.

Conflicts of interests

The authors declare no conflict of interest.

Authors' contribution

Conceptualization, Zhili Chen,Fang Guo and Longmei Luo; Writing—original draft preparation, Zhili Chen, and Fang Guo; Writing—review and editing, Fang Guo; Visualization, Longmei Luo, and Zhili Chen; Supervision, Fang Guo; Funding acquisition, Fang Guo; Resources, Fang Guo; Project administration, Fang Guo. All authors have read and agreed to the published version of the manuscript.

References

[1] Dimitrakopoulos G, Demestichas P. Intelligent transportation systems. *IEEE Veh. Technol. Mag.* 2010, 5(1):77–84.

[2] Adnan MA, Sulaiman N, Zainuddin NI, Besar TBHT. Vehicle speed measurement technique using various speed detection instrumentation. In *2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC)*, Langkawi, Malaysia, 7–9 April 2013, pp. 668–672.

[3] Sun C, Ritchie SG. Individual vehicle speed estimation using single loop inductive waveforms. *J. Transp. Eng.* 1999, 125(6):531–538.

[4] Yu R, Zhang G, Wang Y. Loop detector segmentation error and its impacts on traffic speed estimation. *Transp. Res. Rec.* 2009, 2099(1):50–57.

[5] Ki YK, Baik DK. Model for accurate speed measurement using double-loop detectors. *IEEE Trans. Veh. Technol.* 2006, 55(4):1094–1101.

[6] Klinefelter E, Nanzer JA. Automotive velocity sensing using millimeter-wave interferometric radar. *IEEE Trans. Microwave Theory Tech.* 2020, 69(1):1096–1104.

[7] Adams J. Laser technology for effective and versatile traffic safety systems. *Traffic Technology International '96. Annual Review Issue.* Dorking, Surrey: UKi Media & Events, 1996, pp. 139–141, 143–144 .

[8] Yang Y, Zhang Y, Wang Y, Liu D. Design of 3D Laser Radar Based on Laser Triangulation. *KSII Trans. Internet Inf. Syst.* 2019, 13(5):2414–2433.

[9] Wang Q, Zhu J, Hu F. Ionosphere Total Electron Content Modeling and Multi-Type Differential Code Bias Estimation Using Multi-Mode and Multi-Frequency Global Navigation Satellite System Observations. *Remote Sens.* 2023, 15(18):4607.

[10] Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vancouver, Canada, 24–28 August 1981, pp. 674–679.

[11] Horn BK, Schunck BG. Determining optical flow. *Artif. Intell.* 1981, 17(1–3):185–203.

[12] Yang L, Li Q, Song X, Cai W, Hou C, *et al.* An Improved Stereo Matching Algorithm for Vehicle Speed Measurement System Based on Spatial and Temporal Image Fusion. *Entropy* 2021, 23(7):866.

[13] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 580–587.

[14] Javadi S, Dahl M, Pettersson MI. Vehicle speed measurement model for video-based systems. *Comput. Electr. Eng.* 2019, 76:238–248.

[15] Pelegri J, Alberola J, Llario V. Vehicle detection and car speed monitoring system using GMR magnetic sensors. In *IEEE 2002 28th Annual Conference of the Industrial Electronics Society. IECON 02*, Seville, Spain, 5–8 November 2002, pp. 1693–1695.

[16] Sebastiá JP, Lluch JA, Vizcaíno JRL. Signal conditioning for GMR magnetic sensors: Applied to traffic speed monitoring GMR sensors. *Sens. Actuators, A* 2007, 137(2):230–235.
[17] Zhao Y. Study on Train Speed Measurement Method Based on Multi-source Information Fusion. Ph.D. thesis, Beijing Jiaotong University, 2022.
[18] Qiu TZ, Lu XY, Chow AH, Shladover SE. Estimation of freeway traffic density with loop detector and probe vehicle data. *Transp. Res. Rec.* 2010, 2178(1):21–29.
[19] Iovescu C, Rao S. The fundamentals of millimeter wave sensors. Available: https://www.ti.com/lit/spyy005 (accessed on 27 January 2024).
[20] Du L, Sun Q, Cai C, Bai J, Fan Z, *et al.* A vehicular mobile standard instrument for field verification of traffic speed meters based on dual-antenna Doppler radar sensor. *Sensors* 2018, 18(4):1099.
[21] Du L, Sun Q, Bai J, Wang J. A verification method for traffic radar-based speed meter with target position determination in road vehicle speeding enforcement. *IEEE Trans. Veh. Technol.* 2021, 70(12):12374–12388.
[22] Jeng SL, Chieng WH, Lu HP. Estimating speed using a side-looking single-radar vehicle detector. *IEEE Trans. Intell. Transp. Syst.* 2013, 15(2):607–614.
[23] Bai J, Li S, Zhang H, Huang L, Wang P. Robust target detection and tracking algorithm based on roadside radar and camera. *Sensors* 2021, 21(4):1116.
[24] Jeng SL, Chieng WH, Lu HP. Estimating speed using a side-looking single-radar vehicle detector. *IEEE Trans. Intell. Transp. Syst.* 2013, 15(2):607–614.
[25] Raja Abdullah RSA, Alnaeb A, Ahmad Salah A, Abdul Rashid NE, Sali A, *et al.* Micro-Doppler estimation and analysis of slow moving objects in forward scattering radar system. *Remote Sens.* 2017, 9(7):699.
[26] Roy A, Gale N, Hong L. Automated traffic surveillance using fusion of Doppler radar and video information. *Math. Comput. Modell.* 2011, 54(1–2):531–543.
[27] Liu H, Teng K, Rai L, Zhang Y, Wang S. A two-step abnormal data analysis and processing method for millimetre-wave radar in traffic flow detection applications. *IET Intel. Transport Syst.* 2021, 15(5):671–682.
[28] Bai L, Yang J, Wang J, Lu M. An Overspeed Capture System Based on Radar Speed Measurement and Vehicle Recognition. In *Artificial Intelligence for Communications and Networks: Second EAI International Conference, AICON 2020*. 19–20 December 2020, pp. 447–456.
[29] Zhou YW. Research of multi-sensor integration system for train speed and position measurement. *Appl. Mech. Mater.* 2012, 105:1920–1925.
[30] Du L, Sun Q, Cai C, Bai J, Fan Z, *et al.* A vehicular mobile standard instrument for field verification of traffic speed meters based on dual-antenna Doppler radar sensor. *Sensors* 2018, 18(4):1099.
[31] Zhou Y, Zhou Q, Zheng C, Zhang Q. Rail transit speed measurement method and error analysis based on dual radar. *Control Inf. Technol.* 2021, (01):30–34.
[32] Torres-Guijarro S, Vazquez-Fernandez E, Seoane-Seoane M, Mondaray-Zafrilla JA. A traffic radar verification system based on GPS–Doppler technology. *Measurement* 2010, 43(10):1355–1362.
[33] López AA, de Quevedo AD, Yuste FS, Dekamp JM, Mequiades VA, *et al.* Coherent signal processing for traffic flow measuring radar sensor. *IEEE Sens. J.* 2017, 18(12):4803–4813.
[34] Cho HJ, Tseng MT. A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation. *Math. Comput. Modell.* 2013, 58(1–2):438–448.
[35] Liu Z, Cai Y, Wang H, Chen L, Gao H, *et al.* Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather

conditions. *IEEE Trans. Intell. Transp. Syst.* 2021, 23(7):6640–6653.

[36] Göhring D, Wang M, Schnürmacher M, Ganjineh T. Radar/lidar sensor fusion for car-following on highways. In *The 5th International Conference on Automation, Robotics and Applications*, Wellington, New Zealand, 6–8 December 2011, pp. 407–412.

[37] Zhang J, Xiao W, Coifman B, Mills JP. Vehicle tracking and speed estimation from roadside lidar. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2020, 13:5597–5608.

[38] Göhring D, Wang M, Schnürmacher M, Ganjineh T. Radar/lidar sensor fusion for car-following on highways. In *The 5th International Conference on Automation, Robotics and Applications*, Wellington, New Zealand, 6–8 December 2011, pp. 407–412.

[39] Log MM, Thoresen T, Eitrheim MH, Levin T, Tørset T. Using low-cost radar sensors and action cameras to measure inter-vehicle distances in real-world truck platooning. *Appl. Syst. Innov.* 2023, 6(3):55.

[40] Bonin TA, Choukulkar A, Brewer WA, Sandberg SP, Weickmann AM, *et al.* Evaluation of turbulence measurement techniques from a single Doppler lidar. *Atmos. Meas. Tech.* 2017, 10(8):3021–3039.

[41] Milovanović V. On fundamental operating principles and range-doppler estimation in monolithic frequency-modulated continuous-wave radar sensors. *Facta Univ. Ser.: Electron. Energ.* 2018, 31(4):547–570.

[42] Mandava M, Gammenthaler RS, Hocker SF. Vehicle speed enforcement using absolute speed handheld lidar. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Chicago, IL, USA, 27—30 August 2018, pp. 1–5.

[43] Zhang Y, Hu Q, Xu G, Ma Y, Wan J, *et al.* Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 18953–18962.

[44] Hu Q, Yang B, Xie L, Rosa S, Guo Y, *et al.* Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 14–19 June 2020, pp. 11108–11117.

[45] Li Y, Bu R, Sun M, Wu W, Di X, *et al.* Pointcnn: Convolution on x-transformed points. *Adv. Neural Inf. Process. Syst.* 2018, 31.

[46] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 21–26 July 2017 pp. 652–660.

[47] Yang B, Luo W, Urtasun R. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018, pp. 7652–7660.

[48] Yang B, Liang M, Urtasun R. Hdnet: Exploiting hd maps for 3d object detection. In *Proceedings of The 2nd Conference on Robot Learning*, Zürich, Switzerland, 29–31 October 2018, pp. 146–155.

[49] Li X, Zhang G, Pan H, Wang Z. Cpgnet: Cascade point-grid fusion network for real-time lidar semantic segmentation. In *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 23–27 May 2022, pp. 11117–11123.

[50] Lu T, Ding X, Liu H, Wu G, Wang L. Link: Linear kernel for lidar-based 3d perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 1105–1115.

[51] Koh J, Lee J, Lee Y, Kim J, Choi JW. Mgtanet: Encoding sequential lidar points using long short-term motion-guided temporal attention for 3d object detection. *Proc. AAAI Conf. Artif. Intell.* 2023, 37(1):1179–1187.

[52] Wang H, Shi C, Shi S, Lei M, Wang S, *et al.* Dsvt: Dynamic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 13520–13529.

[53] Ye D, Zhou Z, Chen W, Xie Y, Wang Y, *et al.* Lidarmultinet: Towards a unified multi-task network for lidar perception. 2023, vol. 37 pp. 3231–3240.
[54] Chen Y, Liu J, Zhang X, Qi X, Jia J. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 21674–21683.
[55] Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* 2017, 30.
[56] Yang Z, Sun Y, Liu S, Jia J. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 14–19 June 2020, pp. 11040–11048.
[57] Lang AH, Vora S, Caesar H, Zhou L, Yang J, *et al.* Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, USA, 16–20 June 2019, pp. 12697–12705.
[58] Zeng Y, Hu Y, Liu S, Ye J, Han Y, *et al.* Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving. *IEEE Rob. Autom. Lett.* 2018, 3(4):3434–3440.
[59] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 18–23 June 2018, pp. 4490–4499.
[60] Zhang D, Zheng Z, Niu H, Wang X, Liu X. Fully Sparse Transformer 3D Detector for LiDAR Point Cloud. *IEEE Trans. Geosci. Remote Sens.* 2023, 61:5705212.
[61] Yin T, Zhou X, Krahenbuhl P. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Nashville, TN, USA, 19–25 June 2021, pp. 11784–11793.
[62] Zhan J, Liu T, Li R, Zhang J, Zhang Z, *et al.* Real-aug: Realistic scene synthesis for lidar augmentation in 3d object detection. *arXiv* 2023, arXiv:2305.12853.
[63] Chen Y, Yu Z, Chen Y, Lan S, Anandkumar A, *et al.* Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 1–6 October 2023, pp. 8394–8405.
[64] Dyukov A, Choy S, Silcock D. Accuracy of speed measurements using GNSS in challenging environments. *Asian J. Appl. Sci.* 2015, 3(6).
[65] Wang X, Tu R, Gao Y, Zhang R, Fan L, *et al.* Velocity estimations by combining time-differenced GPS and Doppler observations. *Meas. Sci. Technol.* 2019, 30(12):125003.
[66] Gao Z, Li T, Zhang H, Ge M, Schuh H. Evaluation on real-time dynamic performance of BDS in PPP, RTK, and INS tightly aided modes. *Adv. Space Res.* 2018, 61(9):2393–2405.
[67] Deep A, Mittal M, Mittal V. Application of Kalman filter in GPS position estimation. In *2018 IEEE 8th Power India International Conference (PIICON)*, Kurukshetra, India, 10–12 December 2018, pp. 1–5.
[68] Peyret F, Betaille D, Hintzy G. High-precision application of GPS in the field of real-time equipment positioning. *Autom. Constr.* 2000, 9(3):299–314.
[69] Zhang Z. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece, 20–27 September 1999, pp. 666–673.
[70] Kampelmühler M, Müller MG, Feichtenhofer C. Camera-based vehicle velocity estimation from monocular video. *arXiv* 2018, arXiv:1802.07094.
[71] Schoepflin TN, Dailey DJ. Algorithms for calibrating roadside traffic cameras and estimating mean vehicle speed. In *2007 IEEE Intelligent Transportation Systems Conference*, Seattle, WA, USA, 30 September–3 October 2007, pp. 277–283.
[72] Llorca DF, Salinas C, Jimenez M, Parra I, Morcillo A, *et al.* Two-camera based accurate vehicle speed measurement using average speed at a fixed point. In *2016 IEEE 19th*

International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016, pp. 2533–2538.

[73] Bhardwaj R, Tummala GK, Ramalingam G, Ramjee R, Sinha P. Autocalib: Automatic traffic camera calibration at scale. *ACM Trans. Sens. Netw.* 2018, 14(3-4):1–27.

[74] Javadi S, Dahl M, Pettersson MI. Vehicle speed measurement model for video-based systems. *Comput. Electr. Eng.* 2019, 76:238–248.

[75] Rodríguez-Rangel H, Morales-Rosales LA, Imperial-Rojo R, Roman-Garay MA, Peralta-Peñuñuri GE, *et al.* Analysis of statistical and artificial intelligence algorithms for real-time speed estimation based on vehicle detection with YOLO. *Appl. Sci.* 2022, 12(6):2907.

[76] Lin CJ, Jeng SY, Lioa HW. A real-time vehicle counting, speed estimation, and classification system based on virtual detection zone and YOLO. *Math. Probl. Eng.* 2021, 2021:1–10.

[77] Dahl M, Javadi S. Analytical modeling for a video-based vehicle speed measurement framework. *Sensors* 2019, 20(1):160.

[78] Kamoji S, Koshti D, Dmonte A, George SJ, Pereira CS. Image processing based vehicle identification and speed measurement. In *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 26–28 February 2020, pp. 523–527.

[79] Singla N. Motion detection based on frame difference method. *Int. J. Inf. Comput. Technol.* 2014, 4(15):1559–1565.

[80] Zhan C, Duan X, Xu S, Song Z, Luo M. An improved moving object detection algorithm based on frame difference and edge detection. In *Fourth international conference on image and graphics (ICIG 2007)*, Chengdu, China, 22–24 August 2007, pp. 519–523.

[81] Chen W, Wang W, Wang K, Li Z, Li H, *et al.* Lane departure warning systems and lane line detection methods based on image processing and semantic segmentation: A review. *J. Traffic Transp. Eng. (Engl. ed.)* 2020, 7(6):748–774.

[82] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 2017, 60(6):84–90.

[83] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014, arXiv:1409.1556.

[84] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 7–12 2015, pp. 1–9.

[85] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA , 27–30 June 2016, pp. 770–778.

[86] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, *et al.* Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 2015, 115:211–252.

[87] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 2015, 28.

[88] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, *et al.* End-to-end object detection with transformers. In *European conference on computer vision*. 23–28 August 2020, pp. 213–229.

[89] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30.

[90] Cordonnier JB, Loukas A, Jaggi M. On the relationship between self-attention and convolutional layers. *arXiv* 2019 arXiv:1911.03584.

[91] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* 2020 arXiv:2010.11929.

[92] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, Montreal, QC, Canada, 11–17 October 2021, pp. 10012–10022.
[93] Sun Z, Liu C, Qu H, Xie G. PVformer: pedestrian and vehicle detection algorithm based on Swin transformer in rainy scenes. *Sensors* 2022, 22(15):5667.
[94] Valanarasu JMJ, Yasarla R, Patel VM. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 2353–2363.
[95] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, *et al.* Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 5728–5739.
[96] Liu X, Zhang B, Liu N. CAST-YOLO: An Improved YOLO Based on a Cross-Attention Strategy Transformer for Foggy Weather Adaptive Detection. *Appl. Sci.* 2023, 13(2):1176.
[97] Wang Z, Cun X, Bao J, Zhou W, Liu J, *et al.* Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 17683–17693.
[98] Liang J, Cao J, Sun G, Zhang K, Van Gool L, *et al.* Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, Montreal, QC, Canada, 11–17 October 2021, pp. 1833–1844.
[99] Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, *et al.* Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 3202–3211.
[100] Lian J, Wang D, Zhu S, Wu Y, Li C. Transformer-based attention network for vehicle re-identification. *Electronics* 2022, 11(7):1016.
[101] Zhu X, Su W, Lu L, Li B, Wang X, *et al.* Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* 2020, arXiv:2010.04159.
[102] Dai J, Qi H, Xiong Y, Li Y, Zhang G, *et al.* Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 22–29 October 2017, pp. 764–773.
[103] Meng D, Chen X, Fan Z, Zeng G, Li H, *et al.* Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, Montreal, QC, Canada, 11–17 October 2021, pp. 3651–3660.
[104] Liu S, Li F, Zhang H, Yang X, Qi X, *et al.* Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv* 2022, arXiv:2201.12329.
[105] Li F, Zhang H, Liu S, Guo J, Ni LM, *et al.* Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 13619–13627.
[106] Zhang H, Li F, Liu S, Zhang L, Su H, *et al.* Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* 2022, arXiv:2203.03605.
[107] Zong Z, Song G, Liu Y. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, Paris, France, 1–6 October 2023, pp. 6748–6758.
[108] Zhang S, Chi C, Yao Y, Lei Z, Li SZ. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 14–19 June 2020, pp. 9759–9768.

[109] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 22–29 October 2017, pp. 2980–2988.
[110] Zhao Y, Lv W, Xu S, Wei J, Wang G, *et al.* Detrs beat yolos on real-time object detection. *arXiv* 2023, arXiv:2304.08069.
[111] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, 6–12 September 2014, pp. 818–833.
[112] Lin TY, Maire M, Belongie S, Hays J, Perona P, *et al.* Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, 6–12 September 2014, pp. 740–755.
[113] Sochor J, Juránek R, Španhel J, Maršık L, Širokỳ A, *et al.* Brnocompspeed: Review of traffic camera calibration and comprehensive dataset for monocular speed measurement. *arXiv* 2017, arXiv:1702.06441.
[114] Ribeiro M, Gutoski M, Lazzaretti AE, Lopes HS. One-class classification in images and videos using a convolutional autoencoder with compact embedding. *IEEE Access* 2020, 8:86520–86535.
[115] Russell D. QMUL Junction Dataset. Available: http://www.eecs.qmul.ac.uk/~sgg/QMUL_Junction_Datasets/Junction/Junction.html (accessed on 27 January 2024).
[116] Luvizon DC, Nassu BT, Minetto R. A video-based system for vehicle speed measurement in urban roadways. *IEEE Trans. Intell. Transp. Syst.* 2016, 18(6):1393–1404.
[117] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, Providence, RI, USA, 16–21 June 2012, pp. 3354–3361.
[118] Carreira J, Noland E, Hillier C, Zisserman A. A short note on the kinetics-700 human action dataset. *arXiv* 2019, arXiv:1907.06987.
[119] Flir. FLIR Thermal Starter Dataset Introduction Version 1.3. Available: https://www.flir.com/oem/adas/adas-dataset-form/#anchor29 (accessed on 10 July 2024).
[120] Suo J, Wang T, Zhang X, Chen H, Zhou W, *et al.* HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. *Sci. Data* 2023, 10(1):227.
[121] Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, *et al.* Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 14–19 June 2020, pp. 11621–11631.
[122] Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, *et al.* Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 14–19 June 2020, pp. 2446–2454.
[123] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv* 2015, arXiv:1503.02531.
[124] Fang Y, Sun Q, Wang X, Huang T, Wang X, *et al.* Eva-02: A visual representation for neon genesis. *arXiv* 2023, arXiv:2303.11331.
[125] Fang Y, Wang W, Xie B, Sun Q, Wu L, *et al.* Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 19358–19369.
[126] Wang H, Song K, Fan J, Wang Y, Xie J, *et al.* Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 10375–10385.
[127] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 2020, 33:1877–1901.

[128] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, *et al.* Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, Philadelphia, PA, USA, 18–24 July 2021, pp. 8821–8831.
[129] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 18–24 July 2021, pp. 8748–8763.
[130] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 1–6 October 2023, pp. 4015–4026.
[131] Sofiiuk K, Petrov IA, Konushin A. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 16–19 October 2022, pp. 3141–3145.
[132] Ma J, He Y, Li F, Han L, You C, *et al.* Segment anything in medical images. *Nat. Commun.* 2024, 15(1):654.
[133] Liu Y, Kong L, Cen J, Chen R, Zhang W, *et al.* Segment any point cloud sequences by distilling vision foundation models. *Adv. Neural Inf. Process. Syst.* 2024, 36.
[134] Luo Z, Yan G, Li Y. Calib-anything: Zero-training lidar-camera extrinsic calibration method using segment anything. *arXiv* 2023, arXiv:2306.02656.
[135] Yang H, Ma C, Wen B, Jiang Y, Yuan Z, *et al.* Recognize any regions. *arXiv* 2023, arXiv:2311.01373.
[136] Dhulavvagol PM, Desai A, Ganiger R. Vehical tracking and speed estimation of moving vehicles for traffic surveillance applications. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, Mysore, India, 8–9 September 2017 pp. 373–377.
[137] Biswas D, Su H, Wang C, Stevanovic A. Speed estimation of multiple moving objects from a moving UAV platform. *ISPRS Int. J. Geo-Inf.* 2019, 8(6):259.
[138] Yang L, Li M, Song X, Xiong Z, Hou C, *et al.* Vehicle speed measurement based on binocular stereovision system. *IEEE Access* 2019, 7:106628–106641.
[139] Zhang B, Zhang J. A traffic surveillance system for obtaining comprehensive information of the passing vehicles based on instance segmentation. *IEEE Trans. Intell. Transp. Syst.* 2020, 22(11):7040–7055.
[140] Wang K, Liu M. YOLOv3-MT: A YOLOv3 using multi-target tracking for vehicle visual detection. *Appl. Intell.* 2022, 52(2):2070–2091.
[141] Liu C, Huynh DQ, Sun Y, Reynolds M, Atkinson S. A vision-based pipeline for vehicle counting, speed estimation, and classification. *IEEE Trans. Intell. Transp. Syst.* 2020, 22(12):7547–7560.
[142] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, Beijing, China, 17–20 September 2017, pp. 3645–3649.
[143] Stanojevic VD, Todorovic BT. BoostTrack: boosting the similarity measure and detection confidence for improved multiple object tracking. *Mach. Vision Appl.* 2024, 35(3):1–15.
[144] Ren W, Wang X, Tian J, Tang Y, Chan AB. Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets. *IEEE Trans. Ind. Appl.* 2020, 30:1439–1452.
[145] Chen X, Peng H, Wang D, Lu H, Hu H. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 14572–14581.
[146] Karaev N, Rocco I, Graham B, Neverova N, Vedaldi A, *et al.* Cotracker: It is better to track together. *arXiv* 2023, arXiv:2307.07635.
[147] Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In

2016 IEEE international conference on image processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016, pp. 3464–3468.
[148] Meinhardt T, Kirillov A, Leal-Taixe L, Feichtenhofer C. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 8844–8854.
[149] Sun P, Cao J, Jiang Y, Zhang R, Xie E, *et al.* Transtrack: Multiple object tracking with transformer. *arXiv* 2020, arXiv:2012.15460.
[150] Chu P, Wang J, You Q, Ling H, Liu Z. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, Waikoloa, HI, USA, 2–7 January 2023, pp. 4870–4880.
[151] Wang Z, Wu Y, Niu Q. Multi-sensor fusion in automated driving: A survey. *IEEE Access* 2019, 8:2847–2868.
[152] Yin Y, Zhang J, Guo M, Ning X, Wang Y, *et al.* Sensor fusion of GNSS and IMU data for robust localization via smoothed error state Kalman filter. *Sensors* 2023, 23(7):3676.
[153] Cai H, Zhang Z, Zhou Z, Li Z, Ding W, *et al.* BEVFusion4D: Learning LiDAR-Camera fusion under Bird's-Eye-View via Cross-Modality guidance and temporal aggregation. *arXiv* 2023, arXiv:2303.17099.
[154] Ravindran R, Santora MJ, Jamali MM. Camera, LiDAR, and radar sensor fusion based on Bayesian neural network (CLR-BNN). *IEEE Sens. J.* 2022, 22(7):6964–6974.
[155] Liang T, Xie H, Yu K, Xia Z, Lin Z, *et al.* Bevfusion: A simple and robust lidar-camera fusion framework. *Adv. Neural Inf. Process. Syst.* 2022, 35:10421–10434.
[156] Liu Z, Tang H, Amini A, Yang X, Mao H, *et al.* Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, London, UK, 29 May–2 June 2023, pp. 2774–2781.
[157] Bai X, Hu Z, Zhu X, Huang Q, Chen Y, *et al.* Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 18–24 June 2022, pp. 1090–1099.
[158] Lei K, Chen Z, Jia S, Zhang X. Hvdetfusion: A simple and robust camera-radar fusion framework. *arXiv* 2023, arXiv:2307.11323.
[159] Kim Y, Shin J, Kim S, Lee IJ, Choi JW, *et al.* Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 1–6 October 2023, pp. 17615–17626.
[160] Wang Y, Deng J, Li Y, Hu J, Liu C, *et al.* Bi-LRFusion: Bi-directional LiDAR-radar fusion for 3D dynamic object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18–22 June 2023, pp. 13394–13403.
[161] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 2015, 28.
[162] Sochor J, Juránek R, Herout A. Traffic surveillance camera calibration by 3d model bounding box alignment for accurate vehicle speed measurement. *Comput. Vis. Image Underst.* 2017, 161:87–98.
[163] Gunawan AA, Tanjung DA, Gunawan FE. Detection of vehicle position and speed using camera calibration and image projection methods. *Procedia Comput. Sci.* 2019, 157:255–265.
[164] Ke R, Kim S, Li Z, Wang Y. Motion-vector clustering for traffic speed detection from UAV video. In *2015 IEEE First International Smart Cities Conference (ISC2)*, Guadalajara, Mexico, 25–28 October 2015, pp. 1–5.
[165] Liu Y, Lian Z, Ding J, Guo T. Multiple objects tracking based vehicle speed analysis

with Gaussian filter from drone video. In *Intelligence Science and Big Data Engineering. Visual Data Engineering: 9th International Conference, IScIDE 2019*, Nanjing, China, 17—20 October 2019, pp. 362–373.

[166] Maduro C, Batista K, Peixoto P, Batista J. Estimation of vehicle velocity and traffic intensity using rectified images. In *2008 15th IEEE International Conference on Image Processing*, San Diego, CA, USA, 12–15 October 2008, pp. 777–780.

[167] Doğan S, Temiz MS, Külür S. Real time speed estimation of moving vehicles from side view images from an uncalibrated video camera. *Sensors* 2010, 10(5):4805–4824.

[168] Czajewski W, Iwanowski M. Vision-based vehicle speed measurement method. In *International Conference on Computer Vision and Graphics*, Warsaw, Poland, 20–22 September 2010, pp. 308–315.

[169] Li J, Chen S, Zhang F, Li E, Yang T, *et al.* An adaptive framework for multi-vehicle ground speed estimation in airborne videos. *Remote Sens.* 2019, 11(10):1241.