Data and Metadata. 2025; 4:873
doi: 10.56294/dm2025873

REVIEW

Computer Vision for Vehicle Detection: A Comprehensive Review

Visión artificial para la detección de vehículos: una revisión exhaustive

Soufiane El Asri[1], Khalid ZEBBARA[1], Mohammed AFTATAH[1], Abderrahmane AZAZ[2], Abderrahmane AIT LHOUSSAINE[2], Karim AIT SIDI LAHCEN[2], Mohamed BAARAR[2], Oussama BOUBRINE[2]

[1]IMISR laboratory, Faculty of Applied Sciences Ait Melloul. Agadir, Morocco.
[2]University of Ibn Zohr. Agadir, Morocco.

Cite as: El Asri S, Zebbara K, Aftatah M, Azaz A, AIT LHOUSSAINE A, Ait Sidi Lahcen K, et al. Computer Vision for Vehicle Detection: A Comprehensive Review. Data and Metadata. 2025; 4:873. https://doi.org/10.56294/dm2025873

Submitted: 11-06-2024 **Revised:** 08-10-2024 **Accepted:** 30-03-2025 **Published:** 31-03-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding author: Soufiane El Asri

ABSTRACT

The rapid increase in vehicle numbers has exacerbated challenges in modern transportation, leading to traffic congestion, accidents, and operational inefficiencies. Intelligent Transportation Systems (ITS) leverage computer vision techniques for vehicle detection, improving safety and efficiency. This paper aims to provide a comprehensive review of vehicle detection methods in ITS. Traditional image-processing techniques, including Scale-Invariant Feature Transform (SIFT), Viola-Jones (VJ), and Histogram of Oriented Gradients (HOG), are analyzed. Additionally, modern deep learning-based approaches are examined, distinguishing between two-stage methods such as R-CNN and Fast R-CNN, and one-stage methods like YOLO and SSD. Various image acquisition techniques, including Mono-vision, Stereo-vision, Thermal/Infrared Cameras, and Bird's Eye View, are also reviewed. The analysis highlights the evolution from handcrafted feature-based methods to deep learning techniques, demonstrating significant improvements in detection accuracy and efficiency. One-stage detectors, particularly YOLO and SSD, offer real-time performance, while two-stage methods provide higher precision. The impact of different imaging modalities on detection reliability is also discussed. Advances in deep learning and imaging techniques have significantly enhanced vehicle detection capabilities in ITS. Future research should focus on improving robustness in diverse environmental conditions and optimizing computational efficiency for real-time deployment.

Keywords: Computer Vision; Object Detection; Intelligent Transportation Systems; Vehicle Detection; Deep Learning.

RESUMEN

El rápido aumento en el número de vehículos ha agravado los desafíos en el transporte moderno, causando congestión del tráfico, accidentes e ineficiencias operativas. Los Sistemas de Transporte Inteligente (ITS) utilizan técnicas de visión por computadora para la detección de vehículos, mejorando la seguridad y la eficiencia. Este artículo tiene como objetivo proporcionar una revisión exhaustiva de los métodos de detección de vehículos en ITS. Se analizan técnicas tradicionales de procesamiento de imágenes, incluyendo Transformada de Características Invariantes a la Escala (SIFT), Viola-Jones (VJ) y Histogramas de Gradientes Orientados (HOG). Además, se examinan enfoques modernos basados en aprendizaje profundo, diferenciando entre métodos de dos etapas como R-CNN y Fast R-CNN, y métodos de una etapa como YOLO y SSD. También se revisan diversas técnicas de adquisición de imágenes, como visión monocular, visión estéreo, cámaras térmicas/infrarrojas y vista cenital. El análisis destaca la evolución de los métodos basados en características manuales hacia técnicas de aprendizaje profundo, demostrando mejoras significativas en precisión y eficiencia de detección. Los detectores de una etapa, especialmente YOLO y SSD, ofrecen un rendimiento en tiempo real, mientras que los métodos de dos etapas proporcionan mayor precisión. Asimismo, se analiza el impacto de las diferentes modalidades de imagen en la fiabilidad de la detección. Los avances en aprendizaje profundo y técnicas de imagen han mejorado significativamente la capacidad de detección de vehículos en ITS. Las investigaciones futuras deben centrarse en mejorar la robustez en diversas condiciones ambientales y optimizar la eficiencia computacional para su implementación en tiempo real.

Palabras clave: Visión por Computadora; Detección de Objetos; Sistemas de Transporte Inteligente; Detección de Vehículos; Aprendizaje Profundo.

INTRODUCTION

Intelligent Transportation Systems (ITS) are designed to enhance road safety, optimize traffic flow, and improve transportation efficiency.[1] A fundamental component of ITS is autonomous vehicles, which rely on object detection to perceive and interact with their surroundings. Object detection is a key task in computer vision, enabling vehicles to recognize and track objects such as pedestrians, traffic signs, and other vehicles.

Early object detection methods were based on handcrafted features and traditional image processing techniques, including the Viola-Jones (VJ) algorithm (Viola and Jones, Histogram of Oriented Gradients (HOG),[2] and Scale-Invariant Feature Transform (SIFT).[3] While these methods contributed to initial advancements, they suffered from computational inefficiencies and limited adaptability to complex environments. The introduction of deep learning revolutionized object detection by automating feature extraction and improving accuracy. Models such as R-CNN,[4] Fast R-CNN,[5] Faster R-CNN,[6] YOLO,[7] and SSD enabled real-time detection with high precision.[8] More recently, transformer-based models such as Vision Transformer (ViT), DETR, and Swin Transformer have further enhanced object detection capabilities.

In addition to algorithmic advancements, sensor technologies play a crucial role in vehicle detection. Monovision cameras provide visual data but lack depth perception, which stereovision systems address by offering 3D spatial awareness. LiDAR technology enhances depth accuracy but remains costly. Given the strengths and limitations of these methods, sensor fusion has emerged as a key approach to improving detection robustness in autonomous driving applications.

This review analyzes vehicle detection techniques in ITS, comparing traditional image-processing methods with modern deep learning-based approaches. It also evaluates different sensor acquisition techniques, including monovision, stereovision, and LiDAR, discussing their advantages and limitations. The objective is to assess the effectiveness of these methods and highlight the increasing role of multi-sensor integration in advancing intelligent transportation systems.

METHOD

A systematic bibliographic review was conducted to evaluate the evolution of object detection methods in the context of vehicle detection. The review focused on peer-reviewed articles, conference papers, and technical reports available in prominent databases such as IEEE Xplore, Scopus, Web of Science, and Google Scholar. Keywords used in the search included "object detection," "vehicle detection," "intelligent transportation systems," "autonomous vehicles," "deep learning," "sensor fusion," and related terms. The period covered by the review spans from the inception of traditional techniques in the late 1990s through recent advancements up to 2021. Inclusion criteria were established to select studies written in English that provided detailed methodological descriptions and performance analyses. Articles that met these criteria were further processed through a multi-step selection process, which involved an initial screening of titles and abstracts followed by a full-text review to ensure relevance and methodological rigor. The final selection was based on the clarity of contribution and the extent to which the study addressed the evolution of object detection in vehicle detection applications.

RESULTS

The bibliographic review revealed a marked transition from traditional object detection methods based on image processing and handcrafted features to modern deep learning-based approaches. Traditional methods, while foundational, are increasingly supplanted by deep learning techniques that offer enhanced accuracy and speed through end-to-end training paradigms. In particular, the literature shows a pronounced shift toward one-stage detectors, such as YOLO variants, which have become prevalent due to their real-time performance and favorable speed-to-accuracy ratio. Two-stage approaches continue to offer superior precision but are less suited to the stringent temporal demands of autonomous systems. Additionally, the review underscored the growing importance of sensor fusion—combining data from cameras, LiDAR, and other sensors—to overcome the limitations inherent in individual acquisition modalities. This multi-sensor integration is recognized as a critical enabler for improving detection reliability, especially under challenging environmental conditions. These findings justify the need for further investigation into sensor fusion strategies and the continuous refinement of deep learning models to meet the evolving requirements of modern transportation systems.

The increasing research interest in object detection is evident from the rise in publications over the past two decades. As shown in figure 1, the number of studies on this topic has significantly grown, reflecting the rapid advancements and applications of object detection techniques in various domains, including autonomous driving and intelligent transportation systems.

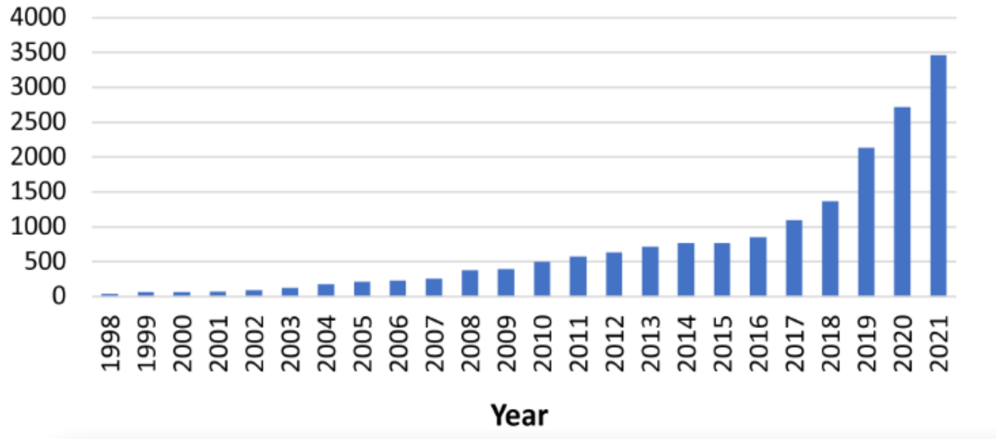

Figure 1. Increasing number of publications in object detection from 1998 to 2021. (Data from Google scholar advanced search: allintitle: "object detection" or "detecting objects".[9]

TRADITIONAL METHODS

Traditional methods can be classified into handcrafted and classical machine learning models.

Scale-Invariant Feature Transform (SIFT)

Lowe introduced Scale-Invariant Feature Transform (SIFT);[3] it is used with a classifier to find objects in images. SIFT is not an object detection method by itself, but it can help by finding special points in an image called key points. These key points are like unique fingerprints of the image and don't change even if the image is rotated, scaled, or lit differently. After SIFT finds these key points, a classifier like SVM, can be used to decide what object these key points belong to.

SIFT features are invariant to image-scaling, translation, rotation and changes in illumination. Meher and Murty used SIFT features to extract the vehicle features of shaded and non-shaded regions of an image.[10]

Viola-Jones detector

Viola and Jones proposed Viola-Jones detector, consisting of a sequence of classifiers.[11] Each classifier is a single perceptron with several binary masks (Haar features). To detect faces in an image, a sliding window is computed over the image. For each image, the classifiers are applied.

VJ uses Haar-like features to capture edges and textures, an integral image for quick calculations, and AdaBoost to select the most important features. A cascade of classifiers speeds up detection by quickly rejecting non-object regions.

Histogram of Oriented Gradients HOG

Dalal and Triggs proposed the Histogram of Oriented Gradients (HOG),[3] a significant improvement over previous methods like SIFT. HOG is a feature descriptor designed to capture object shape and appearance by analyzing edge orientation distributions. This method converts an image of size w×h×3 (width, height, and color channels) into a feature vector of length n, making it suitable for tasks like object detection.

Feature extraction

Feature extraction converts the image into grayscale first, then computes the gradient magnitude and direction by applying simple filters, for instance, the Sobel operator. A gradient representation then divides an image into small cells, where each cell will collect a histogram of edge directions; hence, these histograms are the local shape information in different parts of the image. It then normalizes groups of cells into blocks to make the features more robust against illumination changes, as the gradients can be affected by lighting conditions. Finally, all the histograms are combined into one feature vector representative of the entire image.

https://doi.org/10.56294/dm2025873

HOG + SVM

After feature extraction of HOG, classification is done using the Support Vector Machine (SVM). First, HOG features are extracted from training images, then an SVM is trained that separates different classes, such as human vs. non-human. Once the model is trained, it can classify new images based on their HOG features.

Before the deep learning-based methods dominated object detection, HOG + SVM was widely used. Nowdays, it could be helpful for some use cases, especially real-time, due to efficiency and low computation compared to deep learning models.

Comparison and Conclusion

Traditional object detection algorithms often rely on manual feature extraction, which can be time-consuming, less portable, and error-prone. These methods, such as SIFT, HOG, and Viola-Jones, require domain expertise to design features that are effective for specific tasks. For example, SIFT focuses on scale- invariant keypoints, HOG captures gradient orientations, and Viola-Jones uses Haar- like features. While these techniques were groundbreaking in their time, they have significant limitations like lack of generalization, sensitivity to variations (lightening, viewpoint) and computational costs. Table 1 summarizes some key differences between SIFT, HOG and VJ.

Table 1. Comparison of SIFT, HOG, and Viola-Jones

Feature	**SIFT**	**HOG**	**Viola-Jones (VJ)**
Type	Keypoint-based	Feature descriptor	Cascade classifier
Scale Invariance	Yes	No	No
Rotation Invariance	Yes	No	No
Robust to Illumination	Yes	Moderate	Low
Speed	Slow	Moderate	Fast
Common Usage	Object recognition	Object detection	Face detection
Machine Learning Dependency	Requires classifier	Requires classifier	Uses cascades

DEEP LEARNING METHODS

The advancement in deep learning methods revolutionized the field of object detection.

Two Stage Methods

Two-stage approach have devided the process of identifying and locating objects within an image into two parts, the first step is region proposal where the system looks for potential regions that might contain objects, the second step is object classification which looks at the proposed regions to determine what object exist. The most known two-stage object detection methods includes Faster R- CNN which is proposed by Ren et al.(6) This method has proven its successful detection of vehicles by leveraging the power of CNN. Han proposed a two-stage approach that used stereo vision cues to generate potential object positions and then used extended HOG features and SVM classifiers to verify all hypotheses,(12) enabling the identification of both people and vehicles with high detection accuracy while achieving faster processing speeds.

Qing Luo et al. proposed a multi- scale algorithm for vehicle detection in natural environment, an improvement to the Faster R-CNN.(13) They introduced a feature extraction method using Neural Architecture Search to find the best way to connect different layers, which helps the model to detect small vehicles. The algorithm is tested on UA-DETRAC dataset which is a challenging multi-target dataset for detection and tracking. Compared with other algorithms their method achieved good results in the detection. Faster R-CNN has insufficient ability to detect small vehicles, the average precision for Fast-RCNN is only 14,16 % but their method improved it to 43,64 %.

Djenouri et al. also proposed a two-stage algorithm for vehicle detection, first cleaning the noise from the data using the SIFT extractor,(14) then applying the region proposal approach, as they stated it has the goal of generating bounding boxes for the regions of interests first, using a convolution neural network. after extracting the potential regions that might contain the object, they classified them into vehicles where the refinement of the results is done by regression process. Their second experiment outperformed the other mentioned algorithms in their paper achieving 0,85 of mAP for handling 1,9 annotated vehicles and 200 000 images while the mAP for other algorithms is below 0,75 dealing with the same number of instances.

One Stage Methods

YOLO (You Only Look Once)

YOLO models are a series of one stage object detection models introduced by Redmon et al.,(7) which brought a different approach to object detection by passing the whole image at once through the model, which then predicts both the class probabilities and bounding box coordinates. This process starts by splitting the image into a grid of NxN cells, for each cell a predefined number of bounding boxes is predicted, with their confidence scores and class probabilities, the classes containing the center of an object are responsible for the detection of that object. The improved real-time performance of YOLO models attracted the interest of the research community to apply it in real-time vehicle detection.

Kang et al. proposed YOLO-FA,(15) which aims to reduce the effect of high uncertainty factors like illumination variations, motion blur, occlusion, etc. Using type-1 fuzzy attention (T1FA) where fuzzy entropy is introduced to change the feature map's weights to reduce uncertainty, along with a mixed depth convolution in MetaFormer (MDFormer) used as a token mixer to capture multi-scale perceptual fields in order to detect vehicles of different sizes. their experiments show that T1FA can bring a 3,2 % AP50 boost on the UA- DETRAC vehicle detection dataset, and a 4,2 % and 8,1 % in scenarios of rain and nighttime respectively.

Pan et al. proposed LVD-YOLO,(16) a lightweight vehicle detection model for resource-limited environments, this model uses EfficientNetv2 as its backbone reducing parameters and enhancing feature extraction capabilities, in addition to a bidirectional feature pyramid structure and a dual attention mechanism for efficient information exchange across feature layers improving multiscale feature fusion. Moreover, the model's loss function is refined with SIoU loss to boost regression and prediction performance. Experiments on PASCAL VOC and MS COCO datasets show that this model outperforms YOLOv5s, acheiving a 0,5 % increase in accuracy along with a 64,6 % reduction in FLOPS and a 48,6 % parameters reduction. Liu et al. proposed PV-YOLO,(17) a model based on YOLOv8n for real-time pedestrian and vehicle detection, this implementation focuses on making the model light-weight and efficient. For the architecture, receptive-field attention convolution (RFAConv) are used as the backbone network, for the neck, a bidirectional feature pyramid network (BiFPN) is used instead if the original path aggregation network (PANet) to simplify the feature fusion process. Additionally, a lightweight detection head is introduced to reduce the computational burden and improve the overall detection accuracy. Finally the lightweight C2f module is utilized to compress the model to further decrease the computational costs. Experiments were carried on the BDD100K and KITTI datasets, where PV-YOLO achieved accuracies of 54,5 % and 78,9 % for pedestrian detection and car detection respectively in the BDD100k dataset, which is a 9,5 % and 6,4 % increase from YOLOv8n results, for KITTI, PV-YOLO achieved 88,2 % mAP0.5, a 0,5 % over YOLOv8n.

SSD (Single Shot multibox Detector)

Single Shot multibox Detector are another type of one stage models, similar to YOLO models, it predicts object class and bounding box boundaries in a single forward pass, starting with feature extraction usually using a pretrained model like ResNet or VGG, then its generate multiple scale feature maps to detect objects of different scales, next, a set of pre-defined anchor boxes are generated for each feature map location which represent candidate boxes of various sizes and shapes, then SSD applies convolutions and fully connected operations on each anchor box for classification while at the same time calculating confidence scores of target object existance in each candidate box.

Chen et al. proposed an improved SSD algorithm with MobileNet v2 as the backbone feature extraction network,(18) channel attention mechanism for feature weighting, and a deconvolution module is used to construct a bottom-top feature fusion structure. this algorithm was tested on the BDD100K and KITTI datasets, where it achieved average precision scores of 82,59 % and 84,83 %, respectively.

Zhang and Fu presented a method for traffic density recognition based on deep residual network- single shot multi-box detector (ResNet-SSD) with feature fusion,(19) the deep residual network is used for feature extraction, experiments were conducted on PAS- CAL VOC, MS COCO and custom recorded datasets, on PASCAL VOC, model achieved average precisions of 0,817, 0,819, 0,777 and 0,797 for car, bicycle, bus and motorcycle classes respectively. for MS COCO the model acheived average precision of 0,343 for small-sized traffic objects, and for the custom recorded dataset, the model achieved average precision of 0,913, 0,881, 0,862 and 0,903 for car, motorcycle, bus and truck classes, respectively.

Transformers

Transformers are a type of deep learning architecture that transforms or changes an input sequence into an output sequence. Vaswani et al. introduced in their paper that transformers rely on self-attention mechanisms to process the input.(20) In this context, the paper of Sun et al. proposed an algorithm that detects vehicles based on swin transformer in hazy images,(21) their main features starts with proposing a dehazing model that uses an attention mechanism to dehaze the images using encoding decoding module to extract semantic features from the input hazy image, it starts with passing the image by the encoder first to capture essential features and then through the decoder to reconstruct the image. after that they tried to solve the problem of datasets scarcity in hazy conditions for vehicle detection by collecting and labeling the dataset Haze-Car for training the model and then used Haze-100 dataset for testing it. The next feature is fusing the dehazing module and swin transformer, the first one is reponsible for extracting clean features from hazy images, and the second module is responsible for object classification and localization. After that they compared the algorithm with other object detection algorithms such as YOLO, SSD, Faster-RCNN, etc. The obtained results shows that their accuracy is better than the other algorithms achieving an average precision of 91 % on the Haze-Car dataset and 82,3 % on the Real Haze-100 dataset.

DETR is a fully end-to-end framework for object detection. Unlike traditional methods such as Faster R-CNN, SSD, and YOLO, DETR eliminates the need for predefined region proposals and anchors. It also removes the non-maximum suppression (NMS) step. By leveraging a Transformer-based approach, DETR formulates object detection as an ensemble prediction problem, directly identifying both the object's location and category within an image. This results in enhanced detection speed and accuracy. Zhang et al. introduced TSD-DETR,(22) a lightweight model for traffic sign detection based on RTDETR.

The model achieves a high mean average precision (mAp) of 96,8 % on the Tsinghua-Tencent 100K dataset and 99,4 % on the Changsha University of Science and Technology Chinese Traffic Sign Detection Benchmark dataset, outperforming existing state-of-the-art models

ACQUISITION METHODS

The acquisition phase plays a fundamental role in vehicle detection by providing the necessary data for perception systems. This phase serves as the primary source of environmental information, enabling object detection and tracking in real-world driving conditions. The effectiveness of an autonomous system heavily depends on the accuracy, reliability, and efficiency of these acquisition methods.(23) We distinguish between two primary sensor categories—camera-based systems and LiDAR—each offering unique capabilities for vehicle detection. Camera systems include monovision, using a single camera for object detection and classification, and stereovision, which employs dual cameras to provide depth perception. Additionally, there are other sensors like infrared (IR) and birds-eye cameras. IR cameras excel in detecting heat signatures and are particularly useful in low-light or night-time conditions. Birds-eye cameras provide a panoramic view, aiding in the detection of objects around the vehicle by combining multiple images. In contrast, LiDAR (Light Detection and Ranging) offers high-precision 3D spatial data, providing detailed depth measurements and environmental models. Each of these two categories of technologies—cameras and LiDAR— presents its own advantages and limitations in terms of accuracy, adaptability to various environmental conditions, and computational complexity.

Monovision

Monovision, employing a single camera, is a widely adopted approach in intelligent transportation systems (ITS) due to its cost-effectiveness, simplicity, and ease of integration. This method facilitates vehicle detection, lane tracking, and object recognition by analyzing visual data from a single perspective. However, a fundamental limitation of monovision is its inability to directly capture depth information, making accurate distance estimation inherently challenging. This constraint significantly affects applications such as forward collision avoidance and autonomous navigation, where precise spatial awareness is crucial for optimal performance.(24) To address these challenges, researchers have developed monocular vision-based algorithms that leverage geometric modeling, machine learning, and deep learning techniques to infer depth from a single viewpoint. Methods such as structure-from-motion and supervised depth estimation using convolutional neural networks (CNNs) have demonstrated improved accuracy in depth perception, enhancing the functionality of monovision-based systems.(25) However, these techniques remain susceptible to environmental variations, including lighting changes, occlusions, and adverse weather conditions, which can degrade detection accuracy and system robustness.

Stereovision

Stereovision, inspired by human binocular vision, is a powerful acquisition technique widely used in intelligent systems for depth estimation, 3D reconstruction, and real-time environmental perception. By leveraging two spatially separated cameras, stereovision enables precise distance measurement, making it a key component in autonomous navigation, robotics, and intelligent transportation systems. Recent research highlights its effectiveness in construction automation and vehicle safety systems. In construction machinery, stereovision enhances obstacle detection, terrain adaptability, and autonomous mobility, contributing to safer and more efficient operations.(26) In intelligent vehicles, it plays a crucial role in real-time road recognition and collision prevention, improving driver assistance systems and enhancing road safety compared to monocular vision,(27) stereovision provides superior depth perception without relying on expensive active sensors like LiDAR, making it a coefficient and scalable alternative.

Thermal/Infrared Cameras

Infrared (IR) cameras are essential in vehicle detection, particularly under challenging conditions such as low visibility at night or in foggy environments. By detecting the thermal radiation emitted by objects, IR cameras can identify vehicles and living organisms especially pedestrians by their heat signatures, which makes them a valuable enhancement for intelligent transportation systems (ITS). As part of modern vehicle detection systems, infrared cameras are widely integrated with other sensors like LiDAR and visible light cameras to improve accuracy. Ding et al. demonstrated how IR imaging can aid vehicle and pedestrian detection,[28] particularly in environments where visibility is compromised. Their approach fused normalized grayscale with Felzenszwalb's Histogram of Oriented Gradients (FHOG) to robustly track objects, even under adverse conditions. Further, Sun et al. proposed an enhanced infrared detection model, Multi-YOLOv8, which addresses the inherent challenges of infrared small object detection, such as low pixel size, dim light, and complex backgrounds. Their model uses multi-input data, including optical flow and background suppression images, alongside the original infrared frames to boost detection performance, especially for small, fast-moving objects like vehicles. While IR cameras excel in capturing heat signatures, their limitations include lower resolution compared to visible light cameras, which may hinder fine-detail detection. Environmental conditions such as temperature fluctuations can also affect the thermal readings, potentially impacting the accuracy of vehicle detection. Despite these challenges, integrating IR cameras with deep learning algorithms, such as YOLOv8, and using advanced techniques like attention mechanisms and small-object detection layers, significantly improves the reliability of detection in diverse scenarios.

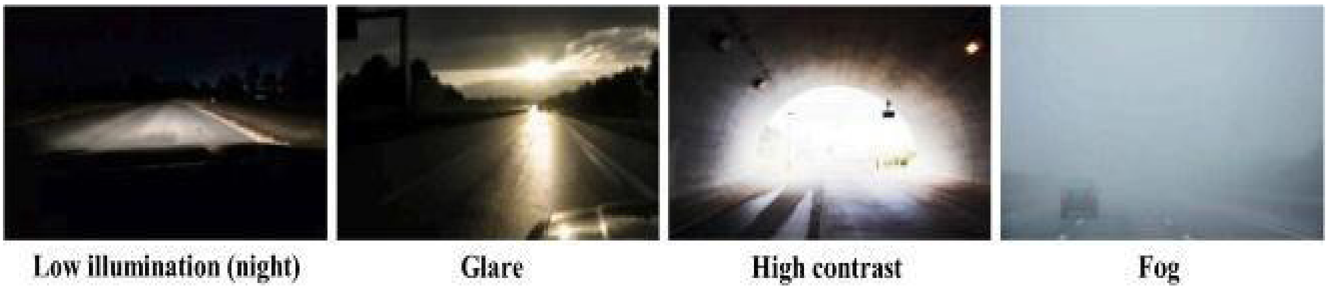

Figure 2. Four typical difficult scenes for a regular RGB camera.[23]

Bird's Eye View

Bird's Eye View (BEV) is increasingly utilized in autonomous driving and intelligent transportation systems (ITS) due to its ability to provide a comprehensive, unobstructed view of the surroundings. The BEV representation aggregates data from various sensors, including cameras, LiDAR, and radar, into a unified top-down perspective. This comprehensive view enhances the detection and tracking of vehicles, pedestrians, and other objects by reducing occlusion and scale issues typically associated with perspective views. The BEV approach is widely used in path planning and collision avoidance systems. Recent research highlights the effectiveness of BEV for real-time vehicle detection. For instance, Wang et al. introduced an efficient and robust multi-camera 3D object detection framework based on BEV.[29] Their method relies on ray-transformation to project multi-camera 2D image features into BEV space, providing a more robust solution to sensor calibration errors. By using a multi-level and multi-scale image encoder and incorporating temporal fusion across multiple frames, their framework significantly improves detection accuracy while maintaining computational efficiency. This approach is particularly advantageous for ITS, where timely and accurate detection of surrounding vehicles is critical for safe navigation. Another significant contribution comes from Eli- jah S. Lee a and Choi b,[30] who proposed a novel method for vehicle localization using BEV, particularly for partially visible vehicles. By incorporating stereo vision, depth information, and geometry, their system accurately estimates the longitudinal and lateral distances of vehicles from a partial appearance, ensuring effective tracking even when only parts of a vehicle are visible (figure 4). The integration of temporal data is another area where BEV methods are evolving. Junhui Zhao emphasizes that by combining multi-source sensor data and temporal information,[31] BEV perception systems can overcome the limitations posed by static single-frame analysis. Temporal fusion improves consistency in object detection and tracking, especially in fast-moving and complex environments. Which proved to be particularly beneficial when monitoring vehicles in dense traffic scenarios. Despite these advancements, challenges remain. BEV perception systems rely heavily on precise camera extrinsic parameters, and errors in calibration or changes in installation can degrade performance. To address this, Wang et al. introduced an extrinsic parameters-free approach,[29] which uses neural networks to predict the necessary parameters online, this improving the robustness of BEV-based vehicle detection systems.

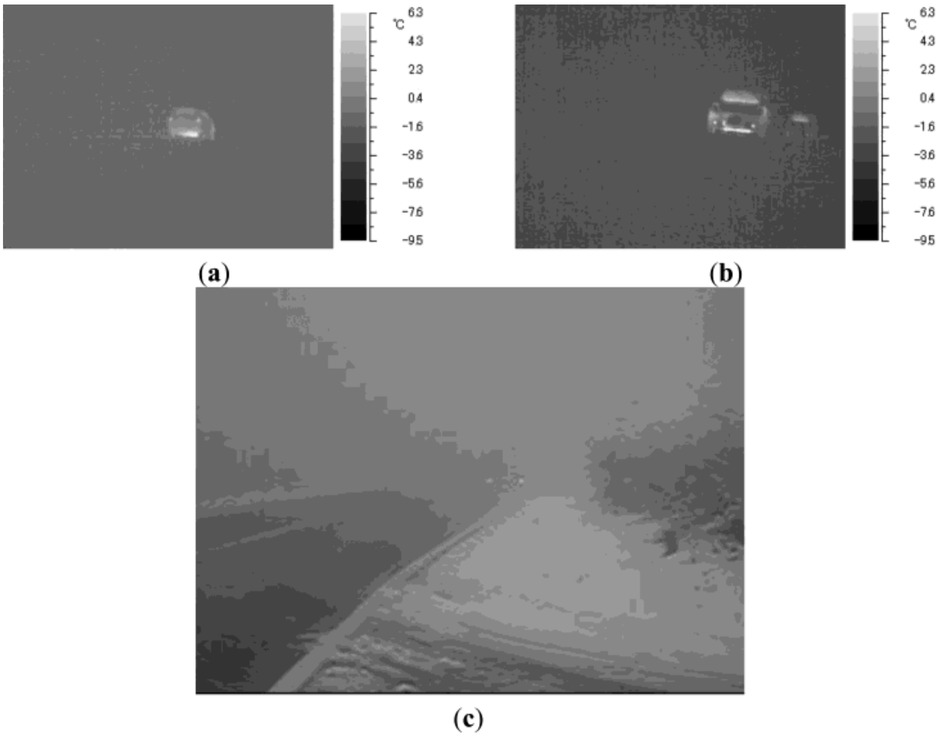

Figure 3. Thermal images and a visible light image in snow and thick fog: (a) thermal image (back side view), (b) thermal image (front side view), and (c) visible light image (front side view).[32]

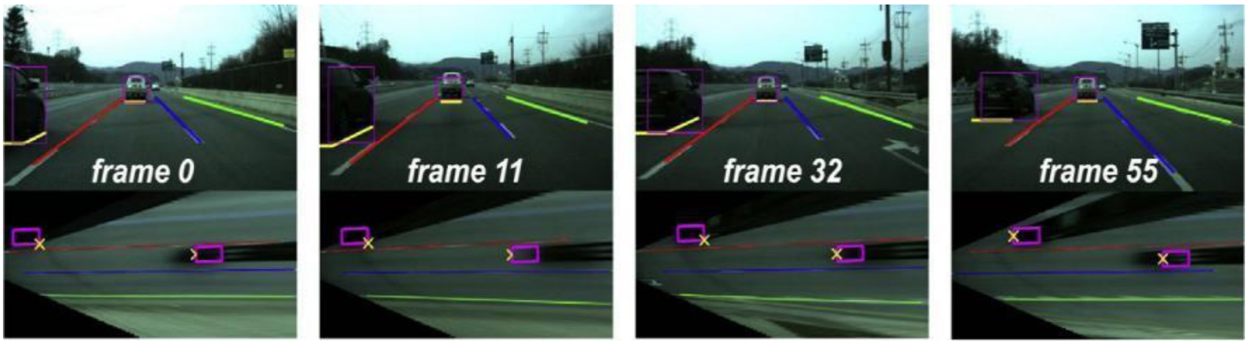

Figure 4. Sequence of vehicle localization results.[30]

LiDAR

Since the development of the laser by Maiman,[33] various LiDAR (Light Detection and Ranging) systems have emerged, which could be valuable for obtaining surface properties that help in object detection and contribute significantly in the field of intelligent driving. LiDAR measures the reflectance of high frequency electromagnetic waves in order to obtain the distance to an object. The commonly used wavelength is 905 nm near-infrared light. which is invisible to humans and also safe due to the absorption. There are 3 ways to generate the laser strating with polygon mirror type which is a spinning mirror with multiple faces to direct a laser beam across a wide range area then a tilt mirror, and the rotating type. One of the challenges that faces this approach is selecting the appropriate type and configuration of the sensors as there are many methods exist.[34] Measuring the time that the light takes to travel from the laser diode to the photodiode can cause latency problems. As mentioned earlier with the increase of LiDAR types each one has different capabilities and characteristics.

The accuracy of the measured distance can vary depending on the sensor type and configuration and also other several factors including the wavelength and environmental conditions. Yang et al. focuses on the difference of ranging accuracy which means considering both the ground truth data and the estimated informations based on the time it takes for the pulses to reflect off objects and return to the sensor in different channels of a vehicle borne LiDAR.[35] Tables 2 and Table 3 shows the results obtained under the conditions of different environments the channel differences of the corrected ranging results are decreased. The decrease rate reached is 82,7-94,8 %.

Table 2. Channel difference comparison of ranging accuracy (unit:m)

Object	Sunny Day	Sunny Night	Rainy Day	Rainy Night
Before	0,2790	0,2708	0,2621	0,2770
After	0,0252	0,0183	0,0136	0,0151
Decrease	90,9 %	93,2 %	94,8 %	94,5 %

Table 3. Channel difference comparison of ranging accuracy (unit:m)

	RMSE			
Object	Sunny day	Sunny night	Rainy day	Rainy night
Before	0,1664	0,1557	0,2229	0,1678
After	0,0149	0,0147	0,0385	0,0286
Decrease	91,0 %	90,5 %	82,7 %	82,9 %

Comparison

In vehicle detection systems, acquisition tools play a significant role in the precision and effectiveness of the overall detection process. The tools in question typically involve different sensor systems such as Monovision, Stereovision, Infrared Cameras, Bird's Eye View (BEV), and LiDAR. Table 4 and Table 5 represent a comprehensive comparison of these acquisition methods based on their key attributes: cost-effectiveness, depth perception, computational complexity, and environmental adaptability.

Table 4. Comparison of vehicle detection acquisition tools

Feature	Monovision	Stereovision	Infrared (IR)	Bird's Eye View (BEV)
Cost	Low	Medium	Medium	High
Environmental Adaptability	Poor in low-light, fog, glare	Better than monovision, good in low-light	Excellent for low-light, limited resolution in harsh weather	Excellent in fog, rain
Depth Perception	None	Excellent (3D depth)	Limited (heat signatures)	Limited (no depth)
Real-Time Processing	High	Moderate	Moderate (sensor fusion needed)	Moderate to High (fusion needed)
Sensor Complexity	Low	Moderate (dual cameras)	High (fusion required)	High (multiple sensors)
Detection Range	Limited	Good (camera separation)	Moderate (thermal-based)	Moderate (wide-angle)

Table 5. Comparison of vehicle detection acquisition tools (LiDAR)

Feature	LiDAR
Cost	High
Environmental Adaptability	*Performs well in diverse weather conditions. Sensitive to rain and fog.*
Depth perception	Excellent (3D depth estimation).
Real-time Processing capability	*Moderate to Low (high computational load).*
Sensor Integration Complexity	High (requires precise calibration).
Detection range	*Very High (long-range and detailed spatial data).*

CONCLUSION

This study examined various computer vision techniques for vehicle detection, highlighting their evolution from traditional methods to deep learning-based approaches. Traditional techniques, such as Haar-like features and Histogram of Oriented Gradients (HOG), rely on handcrafted feature extraction but exhibit limitations in complex environments. In contrast, deep learning methods, particularly one-stage and two-stage detectors, have demonstrated significant advancements. One-stage models like YOLO and SSD offer real-time performance, making them well-suited for transportation systems, while two-stage models prioritize accuracy at the expense of higher computational cost.

A key trend observed in recent research is the preference for one-stage algorithms, particularly YOLO variants, due to their favorable balance between speed and accuracy. Furthermore, the choice of acquisition methods plays a crucial role in detection performance. Monovision cameras enable object recognition but lack depth, whereas stereovision provides 3D perception. Infrared cameras enhance visibility in low-light conditions, while Bird's Eye View (BEV) techniques mitigate occlusions through multiple camera feeds. LiDAR offers precise depth mapping but remains costly. The integration of multiple sensors has emerged as the most effective strategy, enhancing detection accuracy, robustness, and adaptability across diverse environments.

Recent studies increasingly emphasize sensor fusion as the optimal approach, surpassing monovision and LiDAR-based methods in achieving reliable perception for autonomous and assisted driving. This shift underscores the growing recognition of multi-sensor integration as a fundamental component of intelligent transportation systems.

Prospects

In the future, optimizing vehicle detection for embedded and IoT systems will enable real-time processing on low-power devices, making it more efficient for applications like smart surveillance and traffic monitoring. Another key challenge is the detection of similar objects, where vehicles with similar appearances, such as those of the same model or color, may lead to misclassification. Addressing this requires more robust feature extraction and advanced deep learning techniques. Additionally, vehicle congestion remains a critical issue, especially in crowded traffic intersections or parking lots. Enhancing vehicle detection algorithms to perform accurately in dense environments will be essential for improving traffic management and urban mobility solutions. Furthermore, small object detection presents a significant challenge, as distant or partially visible vehicles can be difficult to identify. Developing high resolution feature extraction techniques and multi-scale detection approaches will be crucial to improving detection accuracy in such scenarios.

REFERENCES

1. Mohamed Elassy, Mohammed Al-Hattab, Maen Takruri, and Sufian Badawi. Intelligent transportation systems for sustainable smart cities. Transportation Engineering, 16:100252, 2024. ISSN 2666-691X. doi: https://doi.org/10.1016/j.treng.2024.100252.

2. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

3. David G Lowe. Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision, volume 2, pages 1150–1157. Ieee, 1999.

4. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580– 587, 2014.

5. Ross Girshick. Fast r-cnn. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 1440-1448, 2015. doi: 10.1109/ICCV.2015.169.

6. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1137-1149, 2017. doi: https://doi.org/10.1109/TPAMI.2016.2577031.

7. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, realtime object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

8. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016.

9. Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. CoRR, abs/1905.05055, 2019. URL http://arxiv.org/abs/1905.05055.

10. Saroj K. Meher and M.N. Murty. Efficient method of moving shadow detection and vehicle classification.

AEU-International Journal of Electronics and Communications, 67(8):665–670, 2013. ISSN 1434-8411.doi: https://doi.org/10.1016/j.aeue.2013.02.001.

11. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990517.

12. F. HAN. A two-stage approach to people and vehicle detection with hog-based svm. Proc. of Workshop on Performance Metrics for Intelligent Systems, 2006, 2006.

13. Ji qing Luo, Hu sheng Fang, Fa ming Shao, Yue Zhong, and Xia Hua. Multi-scale traffic vehicle detection based on faster r-cnn with nas optimization and feature enrichment. Defence Technology, 17(4):1542–1554, 2021. ISSN 2214-9147. doi: https://doi.org/10.1016/j.dt.2020.10.006.

14. Youcef Djenouri, Asma Belhadi, Gautam Srivastava, Djamel Djenouri, and Jerry Chun-Wei Lin. Vehicle detection using improved region convolution neural network for accident prevention in smart roads. Pattern Recognition Letters, 158:42–47, 2022. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2022.04.012.

15. Li Kang, Zhiwei Lu, Lingyu Meng, and Zhijian Gao. Yolo-fa: Type-1 fuzzy attention based yolo detector for vehicle detection. Expert Systems with Applications, 237:121209, 2024. ISSN 0957-4174.doi: https://doi.org/10.1016/j.eswa.2023.121209.

16. Hao Pan, Shaopeng Guan, and Xiaoyan Zhao. Lvd-yolo: An efficient lightweight vehicle detection model for intelligent transportation systems. Image and Vision Computing, 151:105276, 2024. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2024.105276.

17. Yuhang Liu, Zhenghua Huang, Qiong Song, and Kun Bai. Pv-yolo: A lightweight pedestrian and vehicle detection model based on improved yolov8. Digital Signal Processing, 156:104857, 2025. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2024.104857.

18. Zhichao Chen, Haoqi Guo, Jie Yang, Haining Jiao, Zhicheng Feng, Lifang Chen, and Tao Gao. Fast vehicle detection algorithm in traffic scene based on improved ssd. Measurement, 201: 111655, 2022. ISSN 0263-2241. doi: https://doi.org/10.1016/j.measurement.2022.111655.

19. Qiang Zhang and Yuguang Fu. Effective traffic density recognition based on resnet-ssd with feature fusion and attention mechanism in normal intersection scenes. Expert Systems with Applications, 261:125508, 2025. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.125508.

20. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

21. Zaiming Sun, Chang'an Liu, Hongquan Qu, and Guangda Xie. A novel effective vehicle detection method based on swin transformer in hazy scenes. Mathematics, 10(13), 2022. ISSN 2227-7390. doi: 10.3390/math10132199.

22. Lili Zhang, Kang Yang, Yucheng Han, Jing Li, Wei Wei, Hongxin Tan, Pei Yu, Ke Zhang, and Xudong Yang. Tsd-detr: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving. Engineering Applications of Artificial Intelligence,139:109536, 2025. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2024.109536.

23. Jingyuan Lei Di Tian, Jiabo Li. Multi-sensor information fusion in internet of vehicles based on deep learning: A review. Neurocomputing, 2025. doi: https://www.sciencedirect.com/science/article/abs/pii/S0925231224016576.

24. Luciano Alonso Rentería Manuel Ibarra Arenado, Juan Maria **Pérez Oria Carlos** TorreFerrero. Monovision-based vehicle detection, distance and relative speed measurement in urban traffic. IET Intelligent Transport Systems, 2013. doi: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-its.2013.0098.

25. Dongsheng Bao and Peikang Wang. Vehicle distance detection based on monocular vision. In 2016

International Conference on Progress in Informatics and Computing (PIC), pages 187– 191, 2016. doi: 10.1109/ PIC.2016.7949492.

26. Tianliang Lin Zhongshen Li Yu Yao Chunhui Zhang Ronghua Ma Zhen Fang, Qihuai Chen, Shengjie Fu, and andHaoling Ren. Automatic walking method of construction machinery based on binocular camera environment perception. Micromachines, 2022. doi: https://doi.org/10.3390/mi13050671.

27. Zidong Han, Junyu Liang, and Jianbang Li. Design of intelligent road recognition and warning system for vehicles based on binocular vision. IEEE Access, 6:62880–62889, 2018. doi: 10.1109/ ACCESS.2018.2876702.

28. Meng Ding, Xu Zhang, Wen-Hua Chen, Li Wei, and Yun-Feng Cao. Thermal infrared pedestrian tracking via fusion of features in driving assistance system of intelligent vehicles. AerospaceEngineering, 2019. doi: https://journals.sagepub.com/doi/10.1177/0954410019890820.

29. Yuanlong Wang, Hengtao Jiang, Guanying Chen, Tong Zhang, Jiaqing Zhou, Zezheng Qing, Chun-yan Wang, and Wanzhong Zhao. Efficient and robust multi-camera 3d object detection in bird-eye-view. Image and Vision Computing, 2025. doi: https://journals.sagepub.com/doi/10.1177/0954410019890820.

30. Dongsuk Kum c Elijah S. Lee a, Wongun Choi b. Bird's eye view localization of surrounding vehicles: Longitudinal and lateral distance estimation with partial appearance. Robotics and Autonomous Systems, 2019. doi: https://journals.sagepub.com/ doi/10.1177/0954410019890820.

31. Jingyue Shi Li Zhuo Junhui Zhao. Bev perception for autonomous driving: State of the art and future perspectives. Expert Systems With Applications, 2024. doi: https://www.sciencedirect.com/science/article/pii/S0957417424019705?via%3Dihub.

32. Masato Misumi andToshiyuki Nakamiya Yoichiro Iwasaki *. Robust vehicle detection under various environmental conditions using an infrared thermal camera and its application to road traffic flow monitoring. sensors, 2013. doi: https://www.mdpi.com/1424-8220/13/6/7756.

33. T.H. Maiman. Stimulated Optical Radiation in Ruby. , 187(4736):493–494, August 1960. doi: https://doi.org/10.1038/187493a0.

34. Jingmeng Zhou. A review of lidar sensor technologies for perception in automated driving. Academic Journal of Science and Technology, 3(3):255–261, Nov. 2022. doi: https://doi.org/10.54097/ajst.v3i3.2993.

35. Tao Yang, Wei Yan, Jiancheng Lai, Yan Zhao, Zhixiang Wu, Yunjing Ji, Chunyong Wang, and Zhenhua Li. Ranging accuracy difference correction among channels of a vehicle-borne road detection lidar. Optics & Laser Technology, 174:110477, 2024. ISSN 0030-3992. doi:https://doi.org/10.1016/j.optlastec.2023.110477. URL: https://www.sciencedirect.com/science/article/pii/S0030399223013701.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Data curation: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Formal analysis: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Research: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Methodology: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Project management: Soufiane El Asri and Khalid ZEBBARA.

Resources: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Software: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Supervision: Khalid ZEBBARA.

Validation: Khalid ZEBBARA.

Display: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Drafting - original draft: Soufiane El Asri, Khalid ZEBBARA, Abderrahmane AZAZ, Abderrahmane AIT LHOUSSAINE, Karim AIT SIDI LAHCEN, Mohamed BAARAR, and Oussama BOUBRINE.

Writing - proofreading and editing: Soufiane El Asri and Mohammed AFTATAH.