Review

Exploring the Unseen: A Survey of Multi-Sensor Fusion and the Role of Explainable AI (XAI) in Autonomous Vehicles

De Jong Yeong *, Krishna Panduru, Joseph Walsh

Posted Date: 20 January 2025

doi: 10.20944/preprints202501.1423.v1

Keywords: autonomous vehicles; self-driving cars; multi-sensor fusion; explainability; explainable artificial intelligence (xai); interpretability; perception; camera; lidar; radar

Review

Exploring the Unseen: A Survey of Multi-Sensor Fusion and the Role of Explainable AI (XAI) in Autonomous Vehicles

De Jong Yeong [1,2,3,*], Krishna Panduru [1,2,3] and Joseph Walsh [1,2,3]

[1] IMaR Research Centre, Munster Technological University, Co. Kerry, V92 CX88, Ireland

[2] School of Science, Technology, Engineering, and Mathematics, Munster Technological University, Co. Kerry, V92 CX88, Ireland

[3] Lero – the Science Foundation Ireland Research Centre for Software, V92 NYD3 Limerick, Ireland

* Correspondence: dejong.yeong@mtu.ie

Abstract: Autonomous vehicles (AVs) rely heavily on multi-sensor fusion to perceive their environment and make critical, real-time decisions by integrating data from various sensors such as radar, cameras, Lidar, and GPS. However, the complexity of these systems often leads to a lack of transparency, posing challenges in terms of safety, accountability, and public trust. This review investigates the intersection of multi-sensor fusion and explainable artificial intelligence (XAI), aiming to address the challenges of implementing accurate and interpretable AV systems. We systematically review cutting-edge multi-sensor fusion techniques, along with various explainability approaches, in the context of AV systems. While multi-sensor fusion technologies have achieved significant advancement in improving AV perception, the lack of transparency and explainability in autonomous decision-making remains a primary challenge. Our findings underscore the necessity of a balanced approach to integrating XAI and multi-sensor fusion in autonomous driving applications, acknowledging the trade-offs between real-time performance and explainability. The key challenges identified span a range of technical, social, ethical, and regulatory aspects. We conclude by underscoring the importance of developing techniques that ensure real-time explainability, specifically in high-stakes applications, to stakeholders without compromising safety and accuracy, as well as outlining future research directions aim at bridging the gap between high-performance multi-sensor fusion and trustworthy explainability in autonomous driving systems.

Keywords: autonomous vehicles; self-driving cars; multi-sensor fusion; explainability; explainable artificial intelligence (xai); interpretability; perception; camera; lidar; radar

1. Introduction

Autonomous vehicles (AVs), also known as self-driving vehicles, are at the forefront of technological innovation with the potential to transform and revolutionize transportation by improving road user safety, efficiency, accessibility, and reducing greenhouse gas emissions [1,2]. At the core of their operation lies the sophisticated capability to perceive, analyze, and respond to highly dynamic and complex driving environments in real time with minimal to no human intervention. AV's perception system relies on the integration of advanced proprioceptive and exteroceptive sensors, robust processing power, complex machine learning (ML) algorithms, and decision-making systems to analyze and interpret complex traffic situations, navigate through unpredictable conditions, and make real-time critical driving decisions autonomously [2]. In our previous research [3], we investigated the architecture of an autonomous driving system from both functional and technical perspectives; highlighting the key components and subsystems that facilitate AVs to operate efficiently based on system design and operational capabilities, specifically in the perception stage of self-driving solutions.

 doi:10.20944/preprints202501.1423.v1

AVs are not limited to on-road applications such as highway driving and navigation or urban driving, nor to off-road environments in industries like agriculture, mining, and construction [4–6]. It extends to a broader range of domains, including maritime settings, where AVs are applied to manage self-navigating vessels, automated container handling and logistic operations in container port terminals, et cetera; hence, improving the safety and efficiency of port activities [7,8]. Whether operating in structured urban settings with well-defined road networks, navigating unstructured and rugged off-road terrains, or coordinating day-to-day logistical tasks within dynamic maritime settings, AVs face diverse operational challenges that demand advanced solutions. All these challenges require efficient and robust multi-sensor fusion and decision-making algorithms to ensure effective and reliable performance.

In AVs, sensors play a pivotal role in perceiving its surroundings and localization of the vehicle within its environment to perform dynamic driving tasks such as obstacle detection and avoidance, path planning, environmental awareness, response to unexpected road situations, et cetera [9,10]. It involves real-time collection and interpretation of large volumes of data (or measurements) from multiple proprioceptive and exteroceptive sensors, including vision cameras, radar, Lidar, ultrasonic sensor, Global Positioning System (GPS), Inertial Measurement Unit (IMU), et cetera. **Table 1** below provides a summary of the commonly adopted proprioceptive and exteroceptive sensors in an AV. It outlines the specific types of sensor that are frequently used in autonomous driving systems to enable robust perception and localization across various operational contexts [11,12].

Table 1. A summary of the commonly utilized proprioceptive and exteroceptive sensors in AVs.

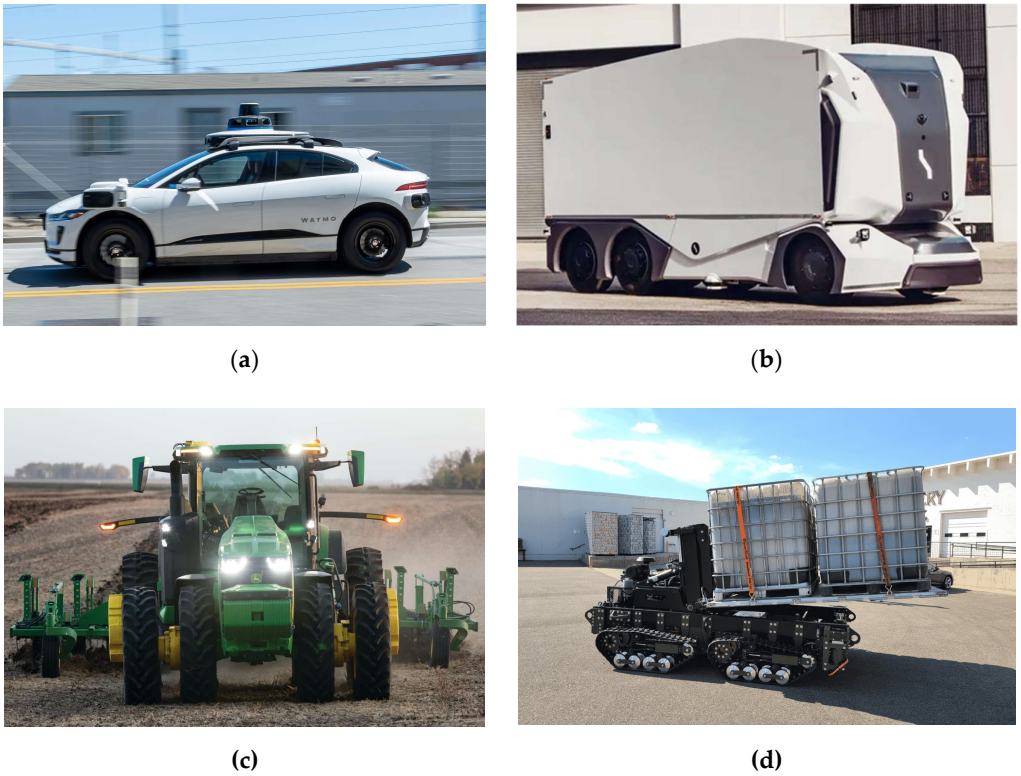
	Definition	**Examples**
Exteroceptive Sensor	It perceives the external environment, detecting objects, obstacles, light intensity, and other relevant features essential for safe navigation.	• Vision cameras. • Radar sensors. • Lidar sensors. • Ultrasonic sensors.
Proprioceptive Sensor	It measures the internal values and gathers information about the dynamic state of a self-driving vehicles, such as its position, speed, and acceleration, that are essential for maintaining stability and ensuring precise control of the vehicle motion.	• IMU. • Global Navigation Satellite System (GNSS). • GPS.

However, the composition of the sensor suite, which refers to the collection of sensors that are integrated into an AV, can vary significantly based on the intended use cases and its specific operational demands. In addition, the specific operational environment of AVs – whether it is on-road, off-road, or in specialized industrial settings – affects the type and arrangement of the sensors that are required to facilitate the perception, localization, and decision-making processes in an autonomous driving system. For example, on-road AVs such as self-driving cars [13] or trucks [14] that operate predominantly on highways and within urban environments often rely heavily on a combination of vision cameras, radar, and Lidars to ensure high-resolution and 360-degree environmental mapping; which are vital in environments where dense traffic and high-speed motion are involved. These sensors must be able to detect and track moving objects, interpret traffic signals, and respond to unpredictable behaviors from other road users.

In contrast, off-road AVs such as autonomous tractor and tillage (agriculture), autonomous pallet loader (military and warehousing), automated rail mounted gantry (RMG) cranes (shipping yards), et cetera [15–17] may employ different sensor configuration that incorporates robustness due to rugged environment, uneven surfaces, low-visibility conditions, or lack of clear infrastructures. In such cases, off-road AVs often incorporate specialized sensors like infrared cameras or thermal cameras to enhance visibility in dusty or low-light conditions [18]. **Figure 1** below presents a visual depiction of various examples of AVs specifically designed for both on-road and off-road applications. The imagery exemplifies the diversity present within the category of AVs, highlighting

 doi:10.20944/preprints202501.1423.v1

how different designs and functionalities are tailored to meet the unique requirements of different operational environments.



(**a**) (**b**)

(**c**) (**d**)

Figure 1. A visual representation of various examples of AVs specifically designed for both on-road and off-road applications. (**a**) Waymo self-driving taxis for ride-sharing services; (**b**) Einride autonomous truck for freight transportation and logistics; (**c**) John Deere autonomous tractor and tillage for agricultural activities and precision farming; (**d**) Stratom autonomous pallet loader for handling pallets. All images shown are provided by the following sources: [14,16,17,19].

The Society of Automation Engineers (SAE) introduced a standardized guideline to eliminate terminological confusion used to describe the varying levels of vehicle automation. It aims to promote clearer communication across industries, enhance risk assessment during system design, support the development of safety and regulatory frameworks, and build public trust and understanding of AV technologies [10,20]. Hence, its initiative has led to the publication of the SAE J3016 standard in 2014, which clearly classifies the levels of driving automation ranging from Level 0 (no automation) to Level 5 (full automation) [21], as illustrated in **Figure 2**. Current automation driving technologies have yet to reach its full potential and have remained at Level 2 (partial automation) for several years [10]. Nonetheless, it is important to highlight that Level 3 (conditional automation) automated driving systems are now being initiated into regular production [22] and some manufacturers, such as Waymo's commercial self-driving ride-sharing services [23], claim to have built vehicles with autonomy that are equivalent to Level 4 (high automation) as described in the SAE J3016 standard. In both on-road and off-road applications, the adoption of this standardized classification supports more coherent development pathways for multi-sensor fusion and explainable artificial intelligence (XAI), as it provides a clearer understanding of the driving system's intended level of autonomy, decision-making responsibilities, and operational limitations.

 doi:10.20944/preprints202501.1423.v1

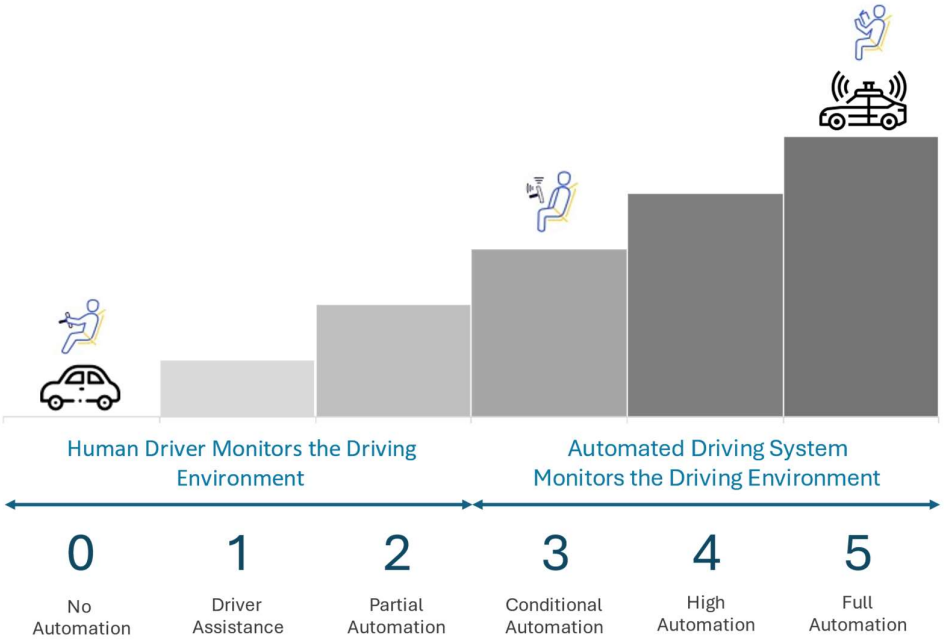


Figure 2. A visual summary of the SAE J3016:2021 standard, which categorizes the levels of driving automation in vehicles. Readers interested in the comprehensive description of the SAE J3016:2021 standard (latest revision) are advised to refer to the SAE International Blog Post [24]. The illustration shown was redrawn and modified based on the diagram in [25,26].

A shared characteristic of an autonomous driving system, applicable to both on-road and off-road applications, is their reliance on multi-sensor fusion, a method that involves integration data from multiple sensor types. This approach is essential for improving the overall perception and situational awareness of AVs, as it helps to address the limitations inherent in individual sensors operating in isolation and mitigate detection uncertainties. For instance, Lidar sensors are highly effective at providing precise, high-resolution depth information, they are susceptible to adverse weather conditions. In contrast, radar sensors are more capable of detecting objects through fog or rain but may offer lower spatial resolution [11]. By integrating data from diverse sensor modalities such as exteroceptive sensors and proprioceptive sensors, multi-sensor fusion significantly enhances the accuracy, reliability, and robustness of the vehicle's perception capabilities. Thus, such an approach enables AVs to achieve a more comprehensive understanding of the surroundings, facilitating more effective navigation in complex and dynamic environments [27,28].

Nonetheless, as the complexity of autonomous driving systems increases, especially with the integration of multiple sensor modalities, the decision-making processes guided by complex deep learning (DL) and ML algorithms often lead to a significant lack of transparency. While these DL and ML models are highly effective at generalizing across a wide range of driving scenarios and are renowned for their powerful ability to model complex patterns through sophisticated data representation, their inner workings and its underlying decision-making logic often results in an inexplainable system [29]. Such systems are concerning in safety-critical applications, such as AVs, where the consequences of erroneous or suboptimal decisions can be severe. For example, in scenarios involving novel conditions or sophisticated driving environments, the inability to understand how or why an autonomous system has made a particular decision can lead to significant risks, including system failures, accidents, or even the loss of human life [30,31]. Hence, it is important to integrate *explainability* into the design of complex autonomous systems to enhance transparency, traceability, accountability, and trust among stakeholders [32].

This paper builds upon and extends the research presented in our previous publication [11], broadening the scope to deliver an in-depth analysis of the intersection between multi-sensor fusion

 doi:10.20944/preprints202501.1423.v1

and XAI in the context of AV systems. In this extended review study, we aim to systematically review state-of-the-art multi-sensor fusion techniques alongside emerging XAI methodologies that contribute to the development of more transparent and interpretable AV systems without compromising safety and perception accuracy. **Section 2** presents an overview of the latest advancements in multi-sensor fusion techniques and provides insight into how multi-sensor fusion methodologies are used to create a unified understanding of the vehicle's surrounding environment. In addition, this section evaluates their respective strengths and weaknesses as well as the challenges associated in real-world autonomous driving applications.

Section 3 outlines the core principles and frameworks of XAI and presents an overview of emerging XAI techniques and tools that can be adopted to enhance the interpretability, transparency, and trustworthiness of an AV system. Besides, this section explores the role of XAI in AVs and emphasizes the critical importance of implementing explainability into the decision-making processes and its challenges to provide clear and interpretable insights into how and why specific driving decisions are made. Lastly, **Section 4** presents a summary overview of the key findings and insights presented throughout the research and highlights future research directions that could contribute to the development of more reliable, interpretable, and trustworthy autonomous driving systems.

2. Multi-Sensor Fusion in Autonomous Vehicles

In AV systems, multi-sensor fusion serves as a cornerstone process in constructing a precise and dependable model of the driving environment. It enables the AV to interpret, predict, and respond to diverse and complex road conditions without little to no human intervention. Unlike traditional vehicles, which rely exclusively on human drivers to perceive and respond to road conditions, AV systems employ a range of sensor types, including cameras, Lidar, radar, and ultrasonic sensors, that capture unique aspects of the driving environment for safe navigations and decision-making [11]. **Figure 3** below provides an illustrative example of a standard sensor configuration for environment perception in AV systems. Nevertheless, it is important to note that the arrangement and integration of various sensors can differ significantly based on the specific application scenarios and operational requirements of the AV [33–37].

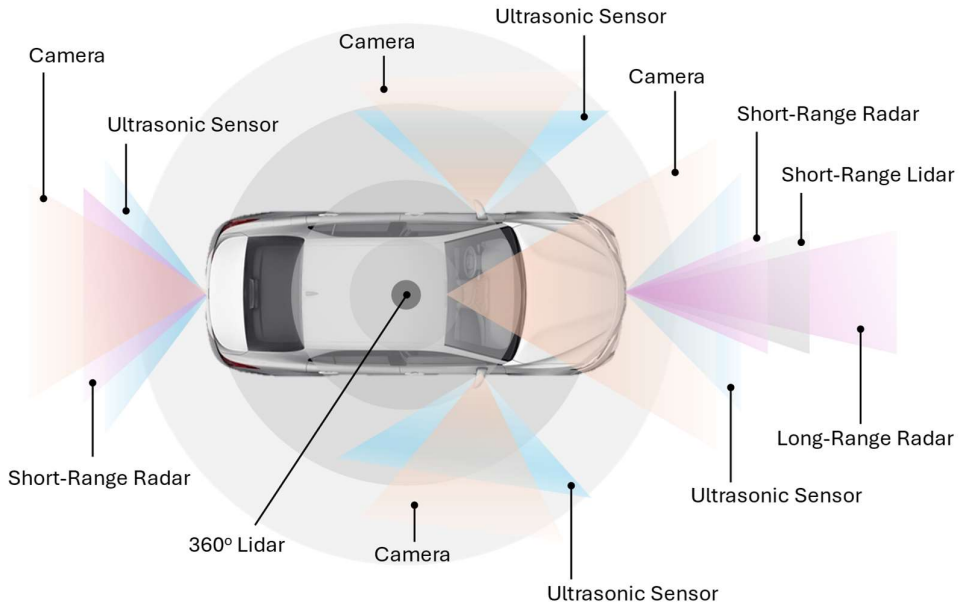


Figure 3. An illustrative example of a typical sensor configuration employed for environmental perception in on-road automated driving systems. It is essential to recognize that the arrangement and integration of sensor

 doi:10.20944/preprints202501.1423.v1

modalities can differ significantly based on operational requirements and specific applications, i.e., off-road versus on-road use cases. Other sensors, such as GPS and IMUs, are not indicated in the illustration. The image shown was redrawn and modified based on the diagram in [36,37].

However, each sensor type carries specific limitations that can compromise its reliability in isolation. For example, cameras deliver high-resolution images that are invaluable for capturing texture and color details and object recognition, but their effectiveness decreases in low light, glare, or adverse weather conditions. Lidar sensors generate detailed depth maps of the surrounding driving environment that enhance spatial awareness, but their performances can degrade under heavy fog or rainy weather conditions [38–40]. Radar sensors, on the other hand, offer reliable distance and velocity measurements without weather condition constraints, but they lack the resolution needed to capture finer details or identify static objects with precision. Lastly, ultrasonic sensors complement the perception suite in AV systems by providing short-range object detection capabilities, which are critical for close-proximity maneuvers such as parking, yet their capabilities are limited in their short operational range and are not suitable for use in high-speed driving scenarios, where higher-resolution data and broader spatial awareness are indispensable [11,41,42]. Therefore, integrating multiple sensor data streams using multi-sensor fusion techniques is imperative for overcoming the limitations that arise when sensors are employed independently. In addition, the multi-sensor fusion process significantly enhances the overall robustness and accuracy of perception in AV systems, which is vital for their performance in dynamic, unpredictable, and safety-critical driving scenarios. **Table 2** below presents a summary of advantages and limitations associated with exteroceptive sensors – cameras, Lidar, radar, and ultrasonic sensors [43,44]. It highlights the strengths and weaknesses of the sensors, offering valuable insights into their performance across different operational requirements and environmental or illumination conditions.

Table 2. An overview of the advantages and limitations associated with exteroceptive sensors: camera, Lidar, radar, and ultrasonic sensors. The table shown is adapted from [44] with modifications.

Exteroceptive Sensors	Advantages	Disadvantages
Camera	• High resolution. • Infrared or thermal sensing available. • Captures texture and color details. • Optimal for object recognition. • Low cost.	• Depth information is not possible without stereo configuration. • Reliant on illumination. • Vulnerable to weather conditions. • Extensive computational power to analyze camera images. • Limited velocity measurements.
Lidar	• Long detection range. • Provides high-resolution three-dimensional (3D) spatial data with distance measurements. • Insusceptible to illumination.	• High cost. • Ineffective and shorter range in adverse heavy rain, fog, or dust. • No texture or color information. • Difficult to detect objects with specular surface or non-Lambertian material [45].
Radar	• Insusceptible to illumination and weather conditions.	• Poor resolutions. • Unable to detect small objects.

 doi:10.20944/preprints202501.1423.v1

	• Offers distance and relative velocity measurements. • Low cost. • Long range.	• Limited classification capability. • Noisy outputs due to reflections. • No texture or color information.
Ultrasonic	• Insusceptible to illumination and weather conditions. • Provides high precision for close-range detection at low speed. • Capable of detecting objects made from all types of materials. • Low cost.	• Limited detection range. • Not suitable for detecting objects at high speed. • Susceptible to interference from wind at high speed. • Sensitive to temperature variation and vapors.

In the context of multi-sensor fusion, several distinct strategies were introduced and adopted to integrate data from multiple sensor modalities to improve the overall perception and decision-making capabilities of AV systems [46]. These strategies can be broadly categorized into three primary approaches: (a) *low-level fusion*, (b) *mid-level fusion*, and (c) *high-level fusion*. Each of these approaches presents a distinct technique for integrating sensor data, designed to optimize the trade-offs between data richness, real-time processing requirements, and computational efficiency. By strategically integrating data at different stages within the sensor data processing pipeline, these fusion techniques aim to address the inherent limitations and uncertainties of individual sensor modalities to create a more robust and resilient perception and navigation model in AV systems. This, in turn, allows AV systems to achieve a higher level of situational awareness, improving the reliability of decision-making and ensuring safer navigation, even in complex and challenging driving environments [11,46–48].

2.1. Multi-Sensor Fusion Approaches

2.1.1. Low-Level Fusion

Low-Level Fusion (LLF), also known as data-level fusion or early fusion [48–50], represents the most granular approach to integrating sensor data in AV systems, where data from multiple sensor types is integrated at the lowest abstraction level, before any significant preprocessing, filtering, or feature extraction occurs. In essence, the LLF approach to multi-sensor fusion utilizes raw features or unprocessed sensor inputs, such as raw radar reflections, camera pixel data, or Lidar point clouds, to create a comprehensive, high-resolution representation of the driving environment. One of the key advantages of LLF approach is its capability to retain the fine-grained information captured by each individual sensor, which maximizes the amount of information available for further analysis including small objects or minute changes in the driving scene. As a result, LLF approach plays an essential role in enhancing the precision and reliability of object detection and environmental awareness in AV's perception system, specifically in dynamic or complex driving scenarios where capturing and preserving fine-grained information is critical for accurate decision-making and ensuring safe navigation [51].

In AV systems, the LLF strategy is often employed in scenarios where high precision and fine-grained detail are indispensable, especially in tasks such as object detection, classification, and tracking. For instance, a recent study by [52] demonstrated that integrating high-resolution camera images and Lidar 3D point clouds at the raw data level substantially improves the accuracy of image depth estimation. It involves projecting Lidar point clouds onto the image plane, otherwise known as sparse depth maps, and further refines into dense depth maps utilizing a depth completion method [53] to transform camera features into a bird's-eye view (BEV) space for long-range high-definition

 doi:10.20944/preprints202501.1423.v1

(HD) map generation; thereby improving the precision of object detection and overall spatial awareness. In addition, the study referenced in [54] introduced a novel camera-radar fusion transformer framework to integrate spatial and contextual information from both the radar and camera sensors using an innovative Spatio-Contextual Fusion Transformer (SCFT) model and a Soft Polar Association (SPA) module. It leverages the complementary strengths of each sensor and the associated polar coordinates between radar points and vision-based object proposals for object detection, classification, and tracking. Such approach achieved state-of-the-art performance on the nuScenes test dataset [55] and outperforming other existing camera-radar fusion methods in terms of accuracy and reliability.

Figure 4 below illustrates the concept and architecture of LLF approach to multi-sensor fusion. It visually demonstrates a high-level overview of the step-by-step fusion processes, emphasizing on how raw data streams from an array of sensor modalities are pre-processed including spatial-temporal calibration [11], prior to being integrated into a unified dataset for further perception and navigation analysis [56,57]. While LLF is advantageous in providing a comprehensive, detailed view of the surrounding environment, it is not without its challenges and drawbacks. LLF requires high computational resources and memory bandwidth to manage and process large volumes of raw data from multiple sensors simultaneously, specifically at high resolutions. It leads to increased latency and may negatively impact the processing capabilities, which are not suitable in complex, dynamic environments where real-time decision-making is essential. Besides, LLF is susceptible to errors in the spatial-temporal calibration of the sensors operating at different frequencies. In safety-critical AV systems, the sensor misalignments can lead to inaccuracies in detecting objects and predicting object distances and trajectories; thus, compromising the reliability and safety of the AV systems. In addition, LLF approach exhibits limited flexibility in scenarios where a sensor fails or malfunctions, as the tightly coupled architecture relies heavily on synchronized inputs from all sensors. Thus, such dependencies reduce the robustness of the system and can pose significant challenges in maintaining the operational safety of the AV system in real-world conditions [56–58].

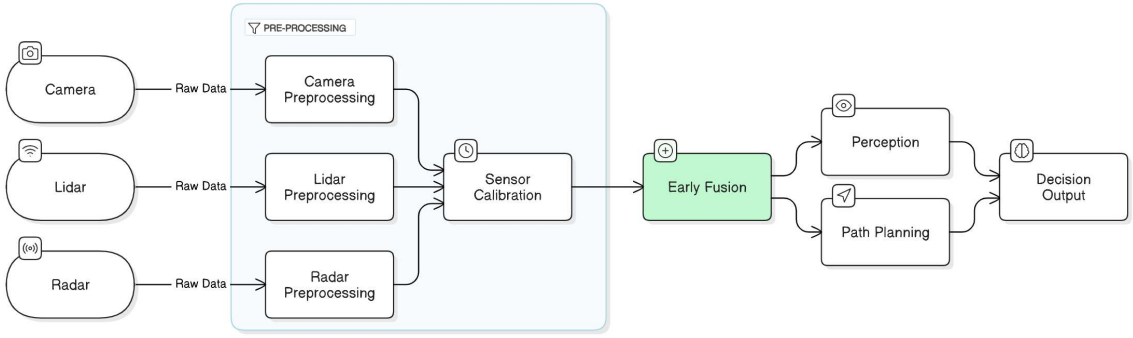


Figure 4. A graphical representation of the concept and architecture of LLF strategy to multi-sensor fusion. It visualizes the step-by-step fusion processes at high level, emphasizing on how raw sensor data streams from multiple sensor modalities are pre-processed, e.g., multi-sensor calibration, prior to being integrated into a unified dataset for further analysis. The diagram illustrated was modified and redrawn based on the depiction in [56,57].

2.1.2. Mid-Level Fusion

In contrast to LLF, which integrates raw data to build a comprehensive and detailed representation of the surrounding driving environment, Mid-Level Fusion (MLF) utilizes the extracted salient features from individual sensor types to construct a more refined and computationally efficient perception of the surroundings. MLF, otherwise known as feature level fusion, intermediate fusion [57], or middle-fusion [59], integrates the high-level features obtained

 doi:10.20944/preprints202501.1423.v1

from individual sensors, such as depth estimations – Lidar, motion trajectories – radar, object boundaries – camera, and et cetera, to develop a more abstract yet informative representation of the environment [48]. MLF approach to multi-sensor fusion lies in its ability to balance perception accuracy with computation efficiency, especially in real-time decision-making scenarios. It offers a pragmatic solution for AV systems by optimizing the allocation of resources and reducing the computational complexity of sensor data processing while maintaining the precision of situational awareness for effective and safe navigation in dynamic, real-world driving conditions [60].

MLF approach is often adopted to achieve a balance between high-accuracy perception and computational efficiency in real-time data processing for object detection, classification, and tracking. In their study, [61] introduced *ContextualFusion*, an environmental-based fusion network, that leverages domain-specific knowledge about the limitations of camera and Lidar sensors, as well as the contextual information about the environment to enhance the perception capabilities. It utilizes the MLF approach to integrate features extracted from the sensors and environmental contextual data, i.e., illumination conditions – daytime and night-time, and rainy weather condition to detect objects in adverse operating conditions, achieving state-of-the-art detection performance on the nuScenes dataset [55] at night-time. In [62], the scholars presented the concept of an end-to-end perception architecture that leverages the MLF strategy in its deep fusion network to create a shared representation of the surroundings. Its fusion network incorporates the features obtained from individual sensor encoders, as well as the temporal dimensions to develop a unified latent space that is sensitive to the nuances of spatial relationships and temporal dynamics for subsequent perception tasks, including object detection, localization, and mapping. By utilizing the unified latent space, the network allows interdependent learning across various perception tasks to minimize redundant data processing; hence, optimizing resource utilization and computational efficiency.

Figure 5 below depicts the concept and architecture of MLF approach to multi-sensor fusion. It illustrates a high-level overview of the sequential fusion processes, emphasizing on how distinct features are initially extracted from individual sensor types prior to being integrated into a shared feature space for subsequent perception and navigation analysis [56,57]. Although MLF offers significant benefits in optimizing resource utilization while maintaining high object detection accuracy, it also presents certain challenges and limitations. MLF requires robust feature extraction algorithms to accurately synthesize the relevant information from disparate sensor sources. It relies on precise feature extraction and is vulnerable to sensor failures, noise, and inconsistencies, which can lead to information loss and resulting in degraded performance in critical perception tasks [48]. Additionally, MLF requires precise multi-sensor spatio-temporal calibration to ensure data consistency during the fusion process. It also requires substantial computational resources to integrate large feature subsets from multiple sensors, which can be challenging in real-time safety-critical systems due to concerns about data latency [11]. Furthermore, as noted in [63], the MLF strategy may not be adequate to support the realization of SAE Level 4 or 5 AVs, as it struggles to handle unexpected scenarios based on predefined feature sets and may fail to retain critical contextual information.

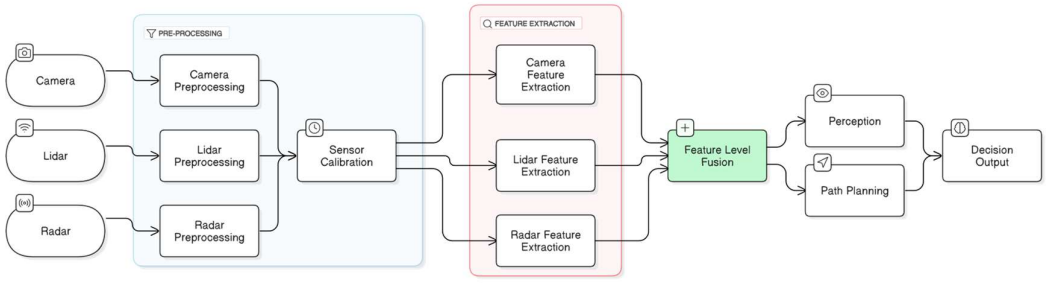


Figure 5. A graphical representation of the concept and architecture of MLF approach to multi-sensor fusion. It visualizes the high-level overview of the MLF processes, where features, such as depth estimations and texture

 doi:10.20944/preprints202501.1423.v1

gradients, were extracted from individual sensors prior to being integrated into a unified dataset for further perception and safe navigation processing to support accurate and safe driving tasks. The diagram shown was redrawn and modified based on the depiction in [57].

2.1.3. High-Level Fusion

High-Level Fusion (HLF), also referred to as decision-level fusion or late fusion [57], represents the highest level of abstraction to integrating multi-sensor data in AV systems. In contrast to LLF and MLF, HLF incorporates individual sensor outputs or decision-making results to construct a comprehensive understanding of the environment. It focuses on integrating the final interpretations or outcomes derived from the analysis performed by individual sensors, such as, location coordinates, velocity vectors, motion trajectories, predicted bounding boxes, classifications of detected objects, et cetera, to establish a reliable, unified, and accurate informed decision [59,64]. One of the key benefits of HLF approach is its modular structure that allows seamless integration of new sensors or updates to existing multi-sensor fusion system without significant changes to the overall fusion framework. As a result, it can be easily adapted to incorporate additional sensing modalities or to accommodate multiple sensor configurations, thereby supporting the scalability of the autonomous driving system [57]. Besides, HLF enhances computational efficiency by focusing on the integration of high-level decisions from individual sensor modalities, which significantly reduces computational complexity compared to raw sensor data, as the processed, abstracted information requires fewer resources, making it beneficial for low latency applications in AV [65]. HLF also promotes robustness and fault tolerance due to its approach to sensor fusion, which allows the system to maintain effective operation when one or more sensors fail or provide erroneous data – no interdependence at the feature or raw data levels.

HLF approach is often adopted to optimize computational efficiency while maintaining effective decision-making capabilities and overall system performance, specifically in real-time, safety-critical applications such as autonomous driving. In their study, [66] introduced a *Multi-modal Multi-class Late Fusion (MMLF)* architecture, which integrates object-level information from various sensor modalities and quantifies the uncertainty associated with the classification results. It involves integrating bounding boxes (spatial locations of objects) from the detectors and a non-zero Intersection over Union (IoU) values to obtain multi-class features for uncertainty estimation. As a result, the integration leads to improved precision and reliability in object detection, achieving substantial performance improvements on the KITTI [67] validation and test datasets. In [68], the researchers presented a late fusion architecture that leverages *Deep Neural Network (DNN)* models to detect pedestrian detection during night-time conditions by utilizing data inputs from RGB and thermal camera images. It involves integrating the outputs, i.e., bounding boxes and detection confidence scores, from individual detection models and applying a *Non-Maximum Suppression (NMS)* method [69] to eliminate redundant detections of the same object and refine the final detection outputs. As a result, the architecture enhances the precision and reliability of pedestrian detection in night-time conditions while ensuring an optimal balance between detection accuracy and low response time during real-time inferencing.

Figure 6 below demonstrates the concept of HLF approach to multi-sensor fusion. It visualizes the high-level overview of the HFL processes, where the outputs generated by individual sensor data analysis are integrated to achieve enhanced situational awareness and reliable informed decisions in dynamic driving scenarios [56,57]. While HLF strategy is advantageous in terms of its computational efficiency and modularity, it is not without its challenges and drawbacks. One notable drawback is the potential loss of detailed contextual information that is often available in raw or feature-level data. HLF may overlook the fine-grained details that are crucial for precise decision-making, especially in dynamic and complex driving environments. The omission of these details can result in erroneous or suboptimal decisions, which can negatively impact the overall performance and safety of the autonomous driving system [59]. Besides, HLF approach relies significantly on the precision and reliability of each individual sensor's interpretation of the surroundings. In other words, any

 doi:10.20944/preprints202501.1423.v1

inaccuracies, misclassifications, or failures in the data from a single sensor can propagate through the AV system, which can lead to misinterpretation of objects or incorrect assessments of driving conditions [48].

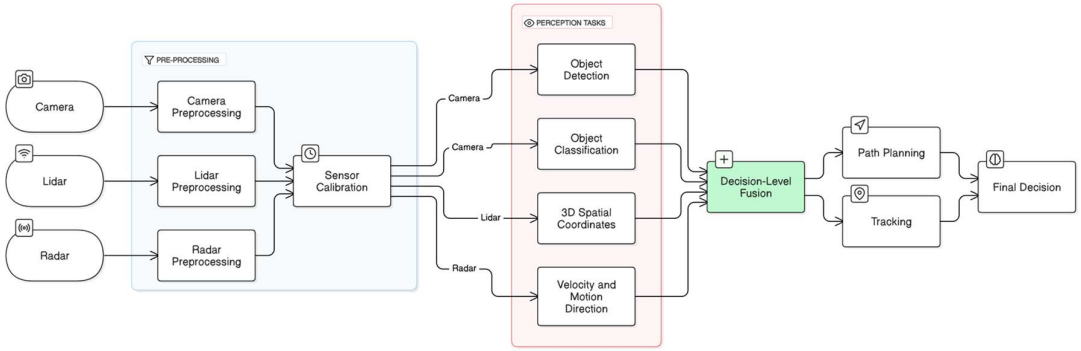


Figure 6. A graphical representation of the conceptual framework of the HLF approach to multi-sensor fusion. It visualizes the high-level overview of the HLF processes, emphasizing on the flow of information as data from individual sensors undertakes independent analysis before the fusion stage occurs to establish a unified informed decision. The depiction shown was adapted and redrawn based on the illustration in [57].

From a computational perspective, sensor fusion can also be categorized into: (a) *centralized fusion*, (b) *decentralized fusion*, and (c) *distributed fusion*. Each of these categories defines the architecture and the specific locus of where the fusion process occurs within the system [70]. In centralized fusion, raw data from each individual sensor is transmitted to a central processing unit, where it is integrated to produce a cohesive and comprehensive representation of the surroundings. In other words, the central processor handles a range of critical tasks in autonomous driving, including data filtering, feature extraction, decision-making, and oversees system control functions, to ensure safe and efficient autonomous driving. In contrast to centralized fusion, decentralized fusion distributes the fusion process across multiple local nodes, where each sensor or subsystem independently processes its data and performs local fusion or analysis before transmitting the processed results to a central unit or other nodes for further integration. In distributed fusion, the concept of decentralization is further extended to allow each sensor or node to share intermediate or partially fusion results across the system without relying on a single central processing unit for final decision-making. **Table 3** below highlights the advantages and drawbacks of centralized fusion, decentralized fusion, and distributed fusion [70–73].

Table 3. An overview of the pros and cons associated with centralized fusion, decentralized fusion, and distributed fusion [70–73].

	Advantages	Disadvantages
Centralized Fusion	• Easy to maintain and update as all data processing occurs in the central processing unit. • High processing power. • Can leverage advanced processing techniques and complex algorithms that require significant computational resources without the need for synchronization across multiple nodes.	• High computational load on the central processor and potentially lead to latency issues. • Single point of failure. • Limited scalability as it can create bottlenecks in both data transmissions and processing power as the number of sensors increases.

 doi:10.20944/preprints202501.1423.v1

Decentralized Fusion	• Efficient multi-sensor data fusion as all data is integrated at a single central processor. • Reduces computational burden on a single processor by distributing processing tasks across multiple nodes. • Robust to failure of individual processing units or one node. • Improves scalability where the system can handle additional sensor modalities without overloading the central processor. • Reduces communication delays and enable faster decision-making by enabling parallel data processing.	• Limited bandwidth especially in high-speed or resource constrained systems. • Complex communication and synchronization can lead to delays or conflicts during data fusion. • Risk of data inconsistency if synchronization is handled ineffectively. • Data redundancy as multiple sensors may perform similar processing tasks independently. • Limited computational resources on individual nodes to process large amounts of data compared to a central processing unit.
Distributed Fusion	• Improves robustness and fault tolerance as the failure of one node or sensor does not compromise the entire system. • Enables faster decision-making as local processing can occur in parallel across different nodes. • Reduces potential bottlenecks and latency. • Flexible and adaptive to changing environments or multi-sensor configurations.	• Requires effective coordination and communication protocols between distributed nodes to ensure seamless integration and synchronization of data. • Increased complexity in data management and fusion due to the distributed nature of the system. • Computational and communication overhead in real-time, large scale, resource-limited systems.

In summary, by strategically integrating sensor data at different stages of the multi-sensor processing pipeline, these multi-sensor fusion approaches aim to leverage the complementary strengths of diverse sensors and the architectural designs of the autonomous driving systems. As discussed, multi-sensor fusion can occur at both the *abstraction* level, i.e., HLF, MLF, and LLF, and *computational* level, i.e., centralized fusion, decentralized fusion, and distributed fusion. On the one hand, the sensor fusion approaches at the abstraction level dictate the timing of when data from individual sensors are integrated. In other words, it addresses the question of "when should the multi-sensor fusion occur?". On the other hand, the fusion approaches at the computational level emphasis on the location of where the fusion process occurs to optimize system performance. In essence, it addresses the question of "where should the multi-sensor fusion occur?". Nonetheless, it is vital for readers to learn that sensor fusion can also occur at the *competition* level, which addresses the question of "what should the fusion do?" [70,72,74] (detailed discussion of the fusion approaches at the competition level, i.e., competitive fusion, coordinated fusion, and complementary fusion is beyond the scope of this manuscript). Ultimately, selecting the most suitable sensor fusion approach

 doi:10.20944/preprints202501.1423.v1

depends on the specific use cases and requirements of the AV systems, including scalability, computational resources, fault tolerance, and real-time performance.

2.2. Fusion Techniques and Algorithms

In AVs, the multi-sensor fusion methods and algorithms serve as the cornerstone for building robust and precise systems that enable reliable perception, accurate localization, and efficient navigation. It supports the integration of data from various sensor types such as GPS, camera, Lidar, and radar sensors, to construct a more comprehensive understanding of the surroundings, thereby, enhancing situational awareness in the highly dynamic and complex driving environment. Over the years, the sensor fusion techniques and algorithms have been studied significantly and well-established in the literature [49,57,75–84]. Fusion techniques and algorithms can be classified into: (a) *traditional* approaches and (b) *advanced* approaches. In traditional approaches, the algorithm utilizes well-established mathematical frameworks, such as deterministic rules, probabilistic theories, and optimization-based criteria, to combine data from multiple sensors. It offers robust, efficient, and interpretable solutions to multi-sensor fusion, specifically in scenarios where the systems require transparency in its decision-making processes and has limited computational resources. Nonetheless, traditional approaches can pose a challenge in nonlinear, highly dynamic, and unstructured environments. Its reliance on predefined models or assumptions about the data distribution may result in suboptimal performance when the assumptions are inaccurate or violated [76].

Conversely, algorithms in advanced approaches leverage complex DL techniques to process, analyze, and integrate data from various sensors. It represents a significant shift towards data-driven methodologies as it employs a multi-layered structure of algorithms (also known as deep neural networks [85,86]) and big data to learn the complex representations, nonlinear relationships, and intricate patterns between multiple sensor inputs for multi-sensor fusion. Essentially, these algorithms are designed to adapt to complex, high-dimensional, and unstructured data, such as camera images, which enables the algorithms to generalize effectively across diverse and dynamic real-world driving environments. As a result, the algorithms provide enhanced perception and navigation capabilities, ensuring reliable performance in challenging and dynamic driving conditions. Nevertheless, as algorithms in advanced approaches continue to advance, their lack of interpretability presents significant challenges in ensuring safety, trust, transparency in its decision-making processes, and accountability, particularly in critical applications such as AV. Besides, DL techniques are computational complex due to its intricate underlying architecture, which can lead to increased latency and resource consumption [11,76,87].

Figures 7 and 8 below demonstrate the traditional and advanced approaches, respectively, highlighting examples of techniques and algorithms that are commonly used in AV systems for tasks such as object detection, localization, and navigation. **Figure 7** exemplifies the traditional fusion algorithms, which include well-established techniques that rely on mathematical models, statistical approaches, knowledge-based theory, and probabilistic frameworks. These techniques are often adopted in scenarios where the dynamics of a system are well understood, and the noise characteristics are predictable [76]. In [88], the scholar utilized the *Unscented Kalman Filter (UKF)* algorithm, an adaptation of the Kalman Filter (KF) algorithm for nonlinear state estimation [89], to incorporate GNSS absolute positioning values and real-time IMU input data. It addresses the potential drift inherent in IMU data during sensor fusion processes, ensuring accurate and reliable estimates of the vehicle's position and orientation and ultimately improving the robustness and precision of the navigation system in AVs. **Figure 8** depicts the advanced fusion algorithms, which leverage modern DL approaches such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Restricted Boltzmann Machine (RBM), Transformers, Reinforcement Learning (RL), and Autoencoders [57,75,90–99]. These techniques are effective in processing complex, high-dimensional input data and are designed to adapt to the dynamic and unpredictable characteristics of real-time driving environments. For example, the scholar in [100] contributed to a novel multi-object tracking system that utilizes three trained *Long Short Term Memory (LSTM)* models to perform

 doi:10.20944/preprints202501.1423.v1

data association, tracking updates, and object position estimation. LSTM model is an RNN-based technique that is designed to capture long-term dependencies in sequential data, which is ideal for tasks like time-series prediction of an object trajectory or vehicle motion prediction [101].

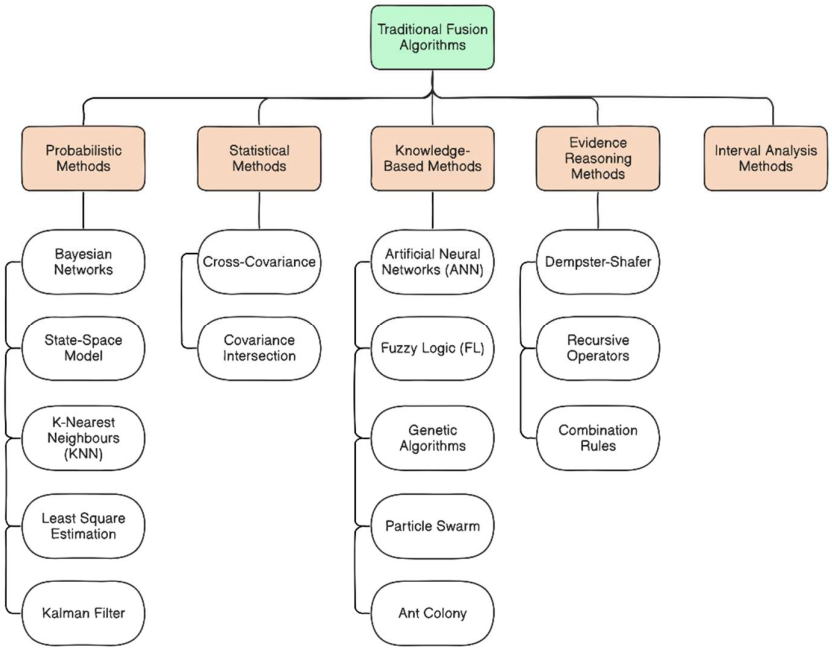
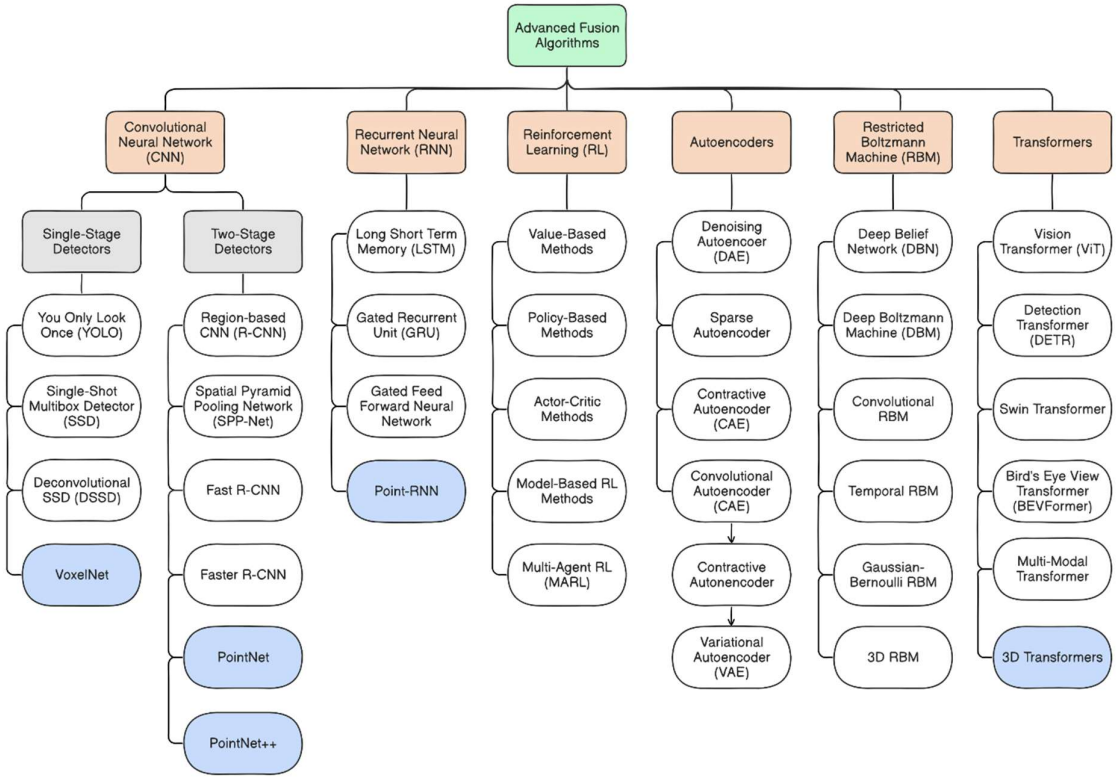


Figure 7. A graphical summary of the traditional fusion methodologies and their associated techniques and algorithms. It highlights the various algorithms used within different paradigms such as probabilistic method, statistical method, knowledge-based method, evidence reasoning method, and interval analysis method. The diagram shown was redrawn based on the illustration in [76].



 doi:10.20944/preprints202501.1423.v1

Figure 8. A graphical overview of the advanced fusion methodologies and their associated techniques and algorithms. It emphasizes the various DL algorithms applied within different paradigms for perception, localization, and mapping systems in AV application. The figure shown was redrawn and adapted based on the depiction in [57,75,76,90–99] to include state-of-the-art algorithms and the algorithms highlighted in "blue" represent those specifically utilized for perception tasks involving 3D point clouds.

In complex applications like autonomous driving systems, traditional and advanced fusion algorithms are commonly utilized in tandem to leverage the strengths of each approach, also known as the hybrid approach [102,103]. This synergistic integration is critical for achieving optimal performance in diverse tasks, such as environmental perception and motion trajectory estimation, where the robustness and efficiency of traditional methods complement the adaptability and learning capabilities of advanced DL algorithms. In [104], the authors proposed a hybrid approach to develop a parameter-free state estimation framework for GPS-based maneuvering-target tracking and localization in AV applications. It features a parameter learning module that integrates a *transformer encoder* architecture with an *LSTM* network to effectively capture the motion characteristics of the system from offline state measurement data. In addition, the framework incorporates the *Expectation-Maximization (EM)* algorithm, which is a well-established statistical approach for parameter estimation in probabilistic models [105]. The EM algorithm estimates the measurement and dynamic characteristics of moving targets in real-time and refines the system parameters based on the outputs of the learning module. Lastly, a *KF* algorithm is used to deliver precise statement estimations, thereby enhancing the accuracy of trajectory tracking predictions. This synergistic integration of traditional algorithms and advanced learning techniques provides a robust solution to estimate state and track trajectory of maneuvering-targets in real time. Hence, it effectively mitigates the impact of sensor noise e.g., Doppler shift, occlusion, and flicker, and eliminates the need to explicitly model the complex dynamics and measurement characteristics of the system.

In [106], the authors introduced YOLO-ACN, a novel and efficient detection framework specifically developed to improve detection precision and overcome the challenges of detecting small targets and occluded objects within complex environments. It includes a lightweight feature extraction network with an attention mechanism, built upon the architecture of the *You Only Look Once (YOLO)* neural network, particularly YOLOv3 [107], to improve focus on small target detection. YOLO is a single-stage detector that simultaneously predicts multiple bounding boxes (detected objects) and class probabilities on an image in real-time [108]. In addition, the network features a modified variant of the NMS classical algorithm, referred to as *Soft-NMS*, within its post-processing phase to eliminate redundant bounding boxes while reducing the likelihood of discarding occluded objects, especially in densely populated environments. Unlike traditional NMS, which eliminates overlapping bounding boxes that exceed the predefined IoU threshold, Soft-NMS retains overlapping boxes with adjusted confidence scores; thereby, improving detection performance in complex scenarios [109,110]. As a result, this synergistic integration has significantly enhanced detection performance and robustness, particularly in recognizing small targets and occluded objects within complex environments, such as urban areas with high pedestrian density.

Ultimately, the selection of the most suitable techniques for the hybrid approach depends on the specific requirements and use cases of the intended application. In complex and dynamic scenarios, leveraging a combination of traditional and advanced algorithms has become a preferred strategy to capitalize on their complementary strengths. This combination not only enhances overall performance but also improves the precision and reliability of the system, ensuring that it is optimized to address the distinct challenges associated with each driving task. **Table 4** below provides an overview of the advantages and weaknesses of both traditional and advanced learning algorithms utilized in multi-sensor fusion systems for AV applications, such as the *UKF*, *Particle Filter (PF)*, *YOLO*, *Dempster-Shafe Theory (DST)*, *PointNet*, and *Faster R-CNN* [11,76,111–129]. Besides, this table focuses on their applications to dynamic driving tasks, such as object detection, tracking, and localization and mapping, which are essential for the safe and efficient operation of autonomous

 doi:10.20944/preprints202501.1423.v1

driving in complex and dynamic driving settings. For a comprehensive discussion of traditional and advanced learning methods for object detection in 3D point cloud data (out of scope in this manuscript), readers are recommended to refer to [57,94,97,130–136].

Table 4. An overview of the advantages and limitations of traditional and advanced learning algorithms employed in multi-sensor fusion systems for AV applications, such as the Unscented Kalman Filter (UKF) algorithm, Particle Filter (PF) algorithm, Dempster-Shafe Theory (DST), YOLO convolutional neural network (CNN), PointNet, and Faster R-CNN.

Algorithms	Descriptions	Applications	Ref.
UKF	UKF is an advanced adaptation of the KF algorithm, specifically developed to address nonlinearities in state estimation with greater efficiency and accuracy. Its strengths and limitations include: • Improved accuracy in nonlinear systems. • Less susceptible to divergence in scenarios where linear approximations might fail. • High computational overhead in high-dimensional systems. • Sensitive to noise modelling. • Requires careful initialization of parameters for optimal performance. • Requires prior knowledge of systems model and data.	• Simultaneous Localization and Mapping (SLAM). • Object tracking.	[111] [112] [115]
Particle Filter (PF)	PF is a recursive algorithm that is utilized to estimate the state of a system by using a set of random samples (particles) to represent the probability distribution, making it ideal for nonlinear and non-Gaussian problems. Its strengths and limitations include: • Highly effective for systems with nonlinear dynamics and non-Gaussian noise. • Scalable for real-time applications with optimization. • Flexible and can integrate data from multiple sensor modalities. • Prone to particle degeneracy. • Sensitive to initial particle distribution, and improper initialization can lead to inaccurate estimates. • High computational cost.	• Object tracking. • Trajectory prediction. • Localization.	[116] [117] [119]
Dempster-Shafer Theory (DST)	DST is a mathematical framework for modeling uncertainties in real-world problems and combining evidence from different sources to make decisions, even if that evidence is uncertain or incomplete, to form a belief about a hypothesis. Its strengths and limitations include:	• Object fusion detection. • Tracking dynamic objects. • Classification. • Decision-making in complex environments.	[113] [120] [121]

 doi:10.20944/preprints202501.1423.v1

	• Does not require pre-defined probabilities. • Integrates evidence from diverse sources with varying reliability. • Improves decision-making by representing varying levels of belief. • Computational expensive in large systems. • Struggles with conflicting evidence. • May produce high uncertainty in complex, high-dimensional data.		
YOLO	YOLO is a real-time object detection algorithm that utilizes a single CNN (single-stage detector) to predict bounding boxes and class probabilities from an image. Several versions of YOLO have been established, each offering improved precision, with the most recent version being YOLOv11 [137]. Its strengths and limitations include: • Fast and able to handle multi-scale object detection in real-time. • Offers high precision in object localization and classification. • Does not require manual feature extraction. • Less accurate than other methods due to coarse bounding boxes. • High computational cost especially in high-resolution images. • Poor detection of occluded objects and small targets.	• Real time object detection. • Traffic sign recognition.	[11] [108] [114] [122]
Faster R-CNN	Faster Region-Convolutional Neural Network (Faster R-CNN) is a two-stage object detection algorithm that utilizes a Region Proposal Network (RPN) and a CNN to detect and localize objects in complex real-world images. Its strengths and limitations include: • High detection precision. • Performs well in cluttered or occluded environments. • Combines region proposal and object classification in a unified framework (end-to-end training). • Requires significant computational resources for training and inference. • Degraded performance when detecting small objects in dense environments. • Slow inference time, which can be challenging for real-time applications.	• Real time object detection.	[76] [114] [123] [124] [125]
PointNet	PointNet is a two-stage detector that introduces a permutation-variant deep neural network to learn global features from	• 3D object detection. • Semantic segmentation.	[126] [127] [128]

 doi:10.20944/preprints202501.1423.v1

	unordered point clouds using a symmetric function, without the need for voxelization. Its strengths and limitations include: • Handles unordered point cloud data. • Can learn directly from raw data without feature engineering. • Sensitive to noisy or sparse data. • Limitations in generalizing to new or unseen scene configurations. • Lack of fine-grained feature extraction but PointNet++ [138] is introduced to address this limitation.	• Localization. • Obstacle detection and avoidance.	[129]

2.3. Challenges in Multi-Sensor Fusion

In AVs, integrating multiple sensor data, otherwise known as multi-sensor fusion, is a cornerstone for implementing precise and robust systems capable of achieving high levels of perception, localization, and mapping essential for autonomous operations. By synergistically integrating information from complementary sensor modalities, multi-sensor fusion allows AVs to construct a comprehensive and dynamic understanding of their environment. In addition, by leveraging unique strengths of various sensors and traditional and advanced fusion algorithms, multi-sensor fusion significantly enhances the capability of AVs to detect obstacles, interpret traffic patterns, and navigate effectively through complex and unpredictable driving environment. Nonetheless, while multi-sensor fusion has revolutionized the capability of AVs to interact effectively with their surroundings, it also introduces several critical technical, operational, and interpretability challenges that need to be addressed for the successful deployment of reliable, safe, scalable, and interpretable (transparent) autonomous systems in real-world applications.

One of the primary challenges is sensor noise, which refers to inaccuracies, inconsistencies, or irrelevant data introduced by individual sensors due to a combination of external interference, hardware limitations, and environmental conditions, such as rain, snow, or dense fog. In [139], the authors presented a comprehensive overview of the challenges associated with radar technologies in autonomous driving systems. A major issue identified is the occurrence of spurious observations, also known as clutter, which arises due to multiple reflections off surfaces in the surroundings, a phenomenon commonly known as multipath. In some cases, such clutter can be difficult to distinguish from real detections, leading to false positive detections in learned radar-based detection models. This, in turn, can significantly undermine the overall system performance and the ability to make precise, reliable, and trustworthy decisions. In our previous exploratory research [11] *(Figure 4)*, we observed multiple instances of false-positive and inconsistent detections within the off-road testing environment, which includes metal objects with corrugated surfaces, traffic cones, and guardrails. These issues were caused by multipath propagation, which distorts sensor signals and leads to inaccurate and unreliable detections in complex environments [140]. A study in [141] showed that Lidar sensors can generate false-positive detections in rainy weather due to reflections from raindrops, and wet surfaces may cause laser beams to scatter, resulting in artifacts such as mirrored objects appearing below the actual ground surface. Therefore, these factors can undermine the accuracy and reliability of the sensor outputs, posing significant challenges for ensuring reliability and precision of autonomous driving operations.

In addition, the heterogeneity of sensor modalities and the ensuing system complexity represent another major challenge in multi-sensor fusion. AVs are generally equipped with a diverse set of sensor types, including cameras, Lidar, radar, ultrasonic sensor, and GPS, each with distinct operational attributes that contribute to their strengths and weaknesses. For example, radar is resilient in poor weather but offers lower spatial resolution; Lidar offers high-resolution depth information but is computationally intensive; and cameras capture rich visual detail but are sensitive

 doi:10.20944/preprints202501.1423.v1

to lighting and weather conditions. Nonetheless, integrating these diverse sensor types introduces significant complexity in algorithmic design and computational processing. It requires sophisticated and innovative fusion algorithms that can handle differences in sensor data format, resolution, and spatial-temporal synchronization [11] while maintaining the overall AV system performance and reliability. Moreover, the complexity of the fusion systems escalates as additional sensors are incorporated to enhance the robustness of perception and support real-time decision-making. It results in the generation of big data, imposing significant demands on computational resources and necessitating innovative real-time processing capabilities to maintain timely and accurate responses. Furthermore, it also intensifies the difficulties associated with testing and validation as rigorous evaluations across varying driving scenarios and environmental conditions are essential to minimize failure risks and ensure dependable and safe operation in real-world contexts [142,143].

In AVs, the volume of data generated by multi-sensor fusion systems is significantly extensive, highlighting the complexity and sophistication of sensor suite employed to perceive and navigate the environment. The continuous operation of these sensors generates high-dimensional, multi-modal data streams, with throughput often reaching multiple gigabytes per second or even terabytes per hour, depending on system configuration (how many sensors are integrated into the system), sensor resolution, refresh rates, and operating conditions [144,145]. This immense data volume is essential for robust perception, localization, and decision-making, but it introduces significant challenges in implementing low-latency data processing pipelines and optimizing the utilization of computational resources. In the event of delays or latency within the data processing pipeline, the AV may fail to respond to dynamic changes in its surroundings, such as unforeseen objects or pedestrians entering the roadway [146]. Besides, the limitations of computational resources in embedded systems that are often utilized in AVs require deliberate trade-offs between accuracy and computational efficiency, needing the optimization of complex fusion algorithms to operate within hardware constraints. Moreover, safety-critical autonomous systems require multi-sensor output verification and cross-validation to address the potential risks of sensor noise, malfunction, or environmental interference; hence, posing significant challenges in its computational load [147]. As a result, addressing these challenges necessitates innovative approaches, such as leveraging parallel processing, hardware accelerators, e.g., Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), and optimized fusion frameworks [148–150].

In addition, multi-sensor fusion systems in AVs are susceptible to malicious attacks, which pose significant risks to the integrity and reliability of their autonomous operation. AVs rely on seamless integration of multiple sensor modalities, but are vulnerable to different forms of adversarial interference, such as spoofing, jamming, and signal manipulation. For example, attackers may broadcast incorrect yet plausible GPS signals to mislead the AV about its true location and leading to navigation inaccuracies [151]. Similarly, adversaries exploit the vulnerabilities of deep neural networks and introduce subtle perturbations to images that are often imperceptible to the human eye, otherwise known as adversarial images. It causes the trained model to produce erroneous predictions or classifications [152]. Moreover, attackers may target the underlying software or communication infrastructure of the multi-sensor fusion system through cyberattacks to overload the system, disrupt data transmission, or manipulate sensor inputs. Thus, these attacks compromise the robustness and reliability of decision-making processes and endanger its overall safety during autonomous operations [153]. In recent years, the *Zero Trust* framework has emerged as a key approach in the design and implementation of multi-sensor fusion systems in AVs. It challenges the traditional assumption of inherent trust within the ecosystem and operates under the core principle that no component or node in the autonomous system should be automatically trusted [154,155]. For a comprehensive exploration of the different attack models and their associated defense strategies (out of scope in this manuscript), readers are encouraged to refer to the research established in [152–154,156–161].

In complex fusion algorithms, the lack of interpretability and explainability presents significant challenges in ensuring transparency and accountability in autonomous operations. One crucial aspect

 doi:10.20944/preprints202501.1423.v1

of this challenge is the necessity to provide clear and comprehensible explanations to stakeholders regarding the decisions and actions made by the autonomous system. For example, end-users often require comprehensible explanations to foster trust and confidence in the reliability of autonomous driving technologies, particularly in safety-critical applications such as AVs. Similarly, regulatory authorities seek comprehensive insights into the decision-making processes to evaluate compliance with well-established safety protocols, legal standards, and ethical guidelines [162]. Additionally, the necessity for explainability is critical for fostering user acceptance of autonomous driving technologies. A lack of clarity in explaining the rationale behind specific actions taken by autonomous systems, especially in situations involving errors or unanticipated outcomes, can significantly undermine user trust and hinder the acceptance of autonomous driving technologies [163–165]. Consequently, overcoming these challenges necessitates a focused effort to design and implement multi-sensor fusion methods and models that strike a balance between complexity and transparency by leveraging XAI techniques to provide valuable insights into how inputs from various sensors are processed and integrated. By enhancing the transparency of decision-making processes, developers can facilitate regulatory approval, enhance confidence and trust among stakeholders, and ensure that autonomous driving systems are reliable and accountable in real-world applications.

3. Explainable Artificial Intelligence (XAI)

XAI, or Explainable Artificial Intelligence, is a specialized domain within the broader discipline of AI that focuses on designing and developing techniques and models that are interpretable and comprehensible to all stakeholders. These stakeholders include, but are not limited to, (a) *researchers* and *academics* aiming to advance the field through theoretical and applied insights; (b) *developers* and *engineers* responsible for developing and maintaining autonomous systems; (c) *end-users* and *consumers* who interact with autonomous systems; (d) *regulators* and *policymakers* to ensure compliance with established standards and safety requirements; and (e) *business leaders* and *industry professionals* focused on utilizing AI to drive commercial and operational success [166–168]. XAI is vital in enhancing transparency, trust, accountability, and safety, especially in safety-critical applications such as autonomous driving. It emphasizes five core principles that serve as foundational pillars, ensuring that such systems conform to transparency, accountability, and user trust standards while achieving their intended functionalities. XAI principles include interpretability, explainability, justifiability, traceability, and transparency, as exemplified in **Figure 9** below [169,170]. It is important for readers to learn that additional XAI principles can encompass fairness, robustness, satisfaction, stability, and responsibility [171] (comprehensive exploration of these principles is beyond the scope of this manuscript).

 doi:10.20944/preprints202501.1423.v1

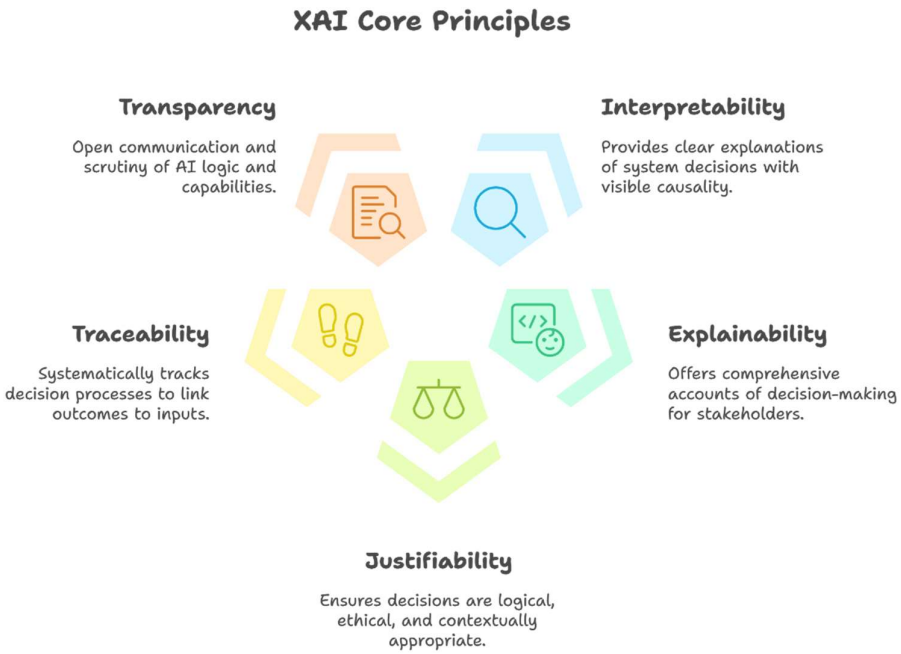


Figure 9. A visual depiction illustrating the five core principles of XAI: interpretability, explainability, justifiability, traceability, and transparency [169,170]. The diagram shown was generated using Napkin AI – an editing platform that transforms text into visual content [172].

- **Interpretability**. It is defined as the ability to explain or to provide clear and comprehensible explanations of the actions and decisions made by the autonomous driving system to relevant stakeholders. It is often deliberated that interpretable systems are more suitable for safety-critical applications, as such systems provide a clear and observable chain of casualties that explains the decision-making processes [173].
- **Explainability**. It is associated with the concept of explanation as a means of providing an interface between humans and a decision-making system that is both an accurate representation of the decision-making process and comprehensive to stakeholders [174]. In essence, explainable systems can provide a clear and detailed account of how and why the decision was made.
- **Justifiability**. It signifies the capability of an artificial intelligence (AI) system to provide logical, ethical, and contextually appropriate reasons for its decisions (outcome) and ensuring alignment with ethical guidelines, user trusts, and accountability [175]. In essence, justifiability ensures that the AI decision made are justifiable and reasonable based on the given data and context. Several approaches can be used to achieve justifiability, including utilizing interpretable models, incorporating post-hoc explanation tools, and involving human experts to review and validate AI decisions [175].
- **Traceability**. It refers to the systematic tracking and documentation of the entire decision-making process of an AI system, ensuring that each action or outcome is traceable to its corresponding inputs, processing steps, reasoning, and outcomes. As a result, any anomalies or errors can be precisely identified and addressed, which is particularly essential in critical situations such as collisions or near-miss events.
- **Transparency**. It involves designing and developing an AI system where the underlying logic, rules, and algorithms governing the decision-making process can be scrutinized and comprehended by all stakeholders. It also involves open and clear communication with stakeholders about the decision-making criteria, functions, capabilities, and limitations of an AI system, e.g., autonomous driving system.

 doi:10.20944/preprints202501.1423.v1

The rapid evolution of ML and DL techniques and algorithms has driven substantial advancements in cutting-edge autonomous applications, such as self-driving vehicles and humanoid robots [176,177]. These advancements underscore the transformative potential of ML and DL technologies in creating systems capable of performing highly sophisticated tasks, such as autonomous driving, with unparalleled precision and efficiency. However, the growing complexity and sophistication of the underlying algorithms pose significant challenges in ensuring transparency and interpretability within complex autonomous systems. In other words, the internal mechanisms of modern ML and DL models, particularly large-scale neural networks, or DNNs, and ensemble methods, are characterized by their opaque nature. Its underlying structure, i.e., multiple hidden layers and extensive parameterization, depicted in **Figure 10** below [178], reflect the difficulties stakeholders encounter in comprehending the internal workings and decision-making processes of these models, resulting in their classification as *black-box* models or systems [179]. Besides, the black-box nature of DNN models introduces additional risks, including the potential propagation of biases and the complexities in diagnosing errors or unintended outcomes. In DNN models, the propagation of biases refers to the amplification or continuation of pre-existing biases embedded in the training data or unintentionally introduced during the design and implementation phases of the DNN models. This issue often arises from imbalances in training datasets, e.g., underrepresentation of specific scenarios, demographic groups, or weather conditions, as well as from implicit assumptions and inconsistencies in labeling practices and feature selection [180]. For example, underlying biases in perception algorithms to detect objects and interpret road signs may lead to disastrous outcomes. As a result, developers use post-hoc analysis techniques to elucidate the decision-making processes of black-box models. However, such methods can be resource intensive, time consuming, and may not always yield definitive explanations, especially when the sources of biases are deeply embedded in complex data or algorithmic structures [171,181–184].

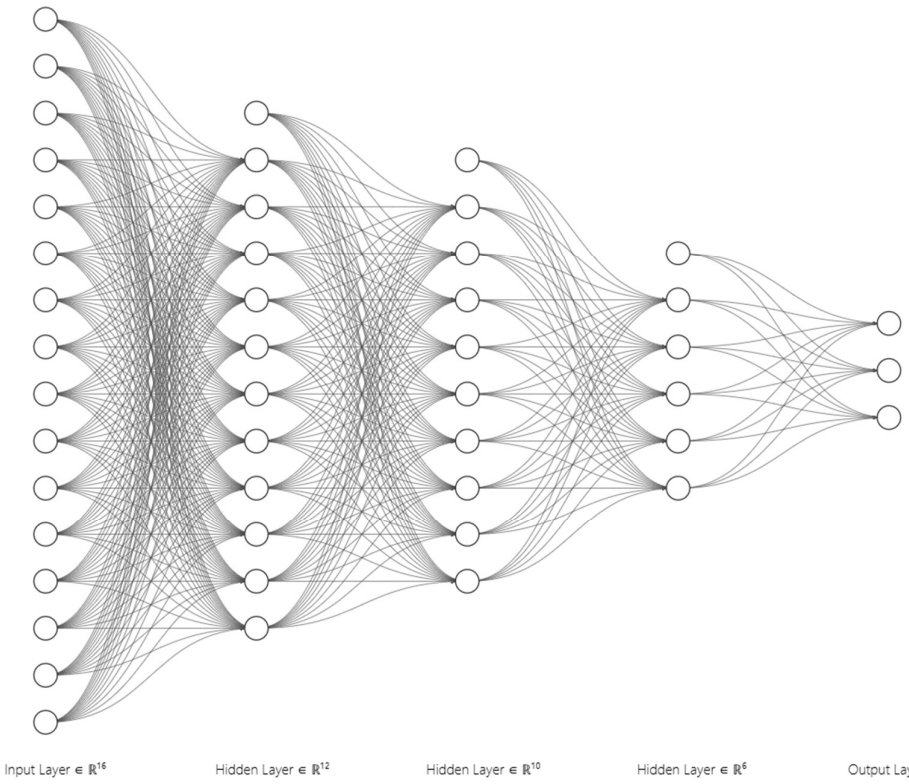


Figure 10. A visual representation of a Deep Neural Network (DNN) model. It shows the underlying architecture of a DNN model, which encompasses an input layer, multiple hidden layers, extensive parameterization, non-

 doi:10.20944/preprints202501.1423.v1

linear activation functions, an output layer, et cetera. The illustration shown is generated using the open-source NN-SVG visualization tool [178].

In contrast to the black-box model, which operates an opaque system with decision-making processes that are difficult to understand, the *white-box* model provides enhanced transparency and offers greater insight into its internal mechanisms. It emphasizes utilizing simple and self-explanatory methods, where the decision-making processes are comprehensible and transparent to human stakeholders. A white-box model is designed with simpler underlying structure and often adopts linear or rule-based traditional algorithms such as, Decision Trees, K-Nearest Neighbors (KNN), Linear Regression, et cetera, which explicitly outline the relationship between inputs and outputs. In linear models, the predicted result can be mathematically expressed as a weighted sum of all its feature inputs, where each feature contributes to the final decision based on its assigned weight [167]. As a result, the white-box model allows a clear and direct understanding and explanation of the decision-making processes. In autonomous driving vehicles, the decision made to decelerate in response to pedestrians crossing the road can be traced and explained through a white-box model. It would generate an audit trail that outlines the rationale behind the action, including factors such as the detection of the pedestrian's location, vehicle's proximity to the pedestrian, and the calculated necessity to decelerate to avoid a potential collision [185]. However, the simplicity and interpretability of white-box models may struggle to attain the same level of predictive accuracy required for handling complex and dynamic real-world autonomous driving tasks, such as object detection. In addition, white-box models are often limited in their ability to effectively handle intricate and unseen scenarios, such as identifying subtle road hazards or reacting to unpredictable driver behavior [167,170,186].

In [169] (Figure 3), the authors presented a comprehensive discussion of the various levels of transparency that represent distinct aspects of interpretability and understanding in ML models. It consists of three distinct levels of transparency: (a) *simulatability*, (b) *decomposability*, and (c) *algorithmic transparency*, which serve as quintessence frameworks for understanding how the internal mechanisms of ML models can be made explainable and accessible to human stakeholders. Within transparency:

- **Simulatability** denotes the ability to simulate the behavior of an ML model through interactive experimentation or human understanding. It enables users to replicate or anticipate the decisions made without necessitating in-depth technical knowledge of its underlying mechanisms or internal architecture. In this aspect, a model is considered simulatable if it can be effectively presented to stakeholders utilizing text, visualizations, or other accessible representations. Furthermore, a simulatable model enables users to reasonably anticipate its outputs based on a given set of inputs, fostering a more intuitive grasp of its decision-making processes [187].
- **Decomposability** refers to the ability to disaggregate an ML model into smaller and interpretable components, such as inputs, parameters, and computations. In essence, decomposability signifies the capability to explain the functioning of a model by examining its constituent elements, providing clarity about how specific inputs influence the outputs, how parameters are optimized, and how intermediate calculations are carried out to reach a final decision. For example, decomposability enables engineers to isolate and explain the contribution of individual subcomponents in autonomous driving, including object detection, trajectory planning, and control systems, which is critical for technical debugging, model refinement, and ensuring compliance with legal and ethical standards. However, in practice, achieving decomposability in intricate ML models, such as DNNs, can be challenging due to their non-linear relationships and the distributed nature of their data representations [169,188].
- **Algorithmic transparency**, as the name suggests, pertains to the extent to which the internal workings and decision-making processes of an algorithm can be clearly understood, elucidated, and scrutinized. In essence, it emphasizes the visibility of how an algorithm operates, from its

 doi:10.20944/preprints202501.1423.v1

initial design through to its decision outputs. In practical terms, algorithmic transparency ensures that the reasoning behind the algorithm decisions can be traced back to its underlying mathematical or computational principles, which are indispensable in identifying and rectifying potential biases, addressing embedded biases, and uncovering unintended behaviors that could compromise the precision and integrity of an ML system. In autonomous driving, understanding the decision-making processes of algorithms, such as how a vehicle decides when to stop or how it identifies and avoids obstacles, is vital in ensuring safety and adherence to regulatory standards. However, the main limitation of algorithmically transparent models is that these models must be fully accessible for analysis using mathematical methods, which is challenging for deep architectures due to the opaque nature of their loss landscapes (multiple interconnected hidden layers) [169,189–192].

The advancement of AI models (ML and DL models) has significantly amplified the need for explainability and interpretability, particularly in safety-critical domains such as autonomous driving. In these domains, it is imperative for AI systems to not only demonstrate high predictive accuracy but also deliver transparent and comprehensible explanations for their decisions to ensure safety, reliability, and adherence to regulatory and ethical guidelines. In XAI, the distinctions between black-box and white-box models underscores a fundamental trade-off in AI models development, i.e., achieving an optimal balance between interpretability and predictive performance. As discussed, black-box models are known for their ability to process complex scenarios with high accuracy but often lack transparency in understanding the underlying processes behind their decision-making. In contrast, white-box models emphasize interpretability and explainability, offering clear and understandable decision-making processes, but may face limitations in managing complex tasks.

However, both paradigms play a pivotal role in addressing the interpretability challenges inherent in cutting-edge, sophisticated AI models, significantly contributing to enhanced accountability and transparency in ML and DL technologies. Besides, both paradigms are instrumental in fostering trust among human stakeholders, which is critical in ensuring the responsible and ethical implementation of autonomous systems within real-world environments. Therefore, addressing interpretability and explainability challenges in autonomous systems has become a primary focus within XAI research, which seeks to develop tools and techniques that can elucidate the decision-making processes of opaque systems and provide human stakeholders with actionable insights into their operations.

3.1. XAI Strategies and Techniques

XAI is an emerging field of research that aims to provide clear, comprehensible, and human-centered explanations for the decisions generated by AI systems. Recent research has investigated several strategies and methodologies designed to elucidate the decision-making processes of intricate and opaque black-box models. XAI methods can be categorized into three main categories: (a) *explanation level*, (b) *implementation level*, and (c) *model dependency* [193]. Such categories offer a systematic framework for understanding the diverse approaches designed to enhance the interpretability and explainability of sophisticated ML and DL systems, especially in contexts where transparency is imperative. It enables researchers and practitioners to select appropriate methods or strategies tailored to specific applications and requirements.

Explanation level refers to the scope and depth of insights delivered, addressing either the overarching behavior of the model or the rationale behind specific individual instances. This concept is subdivided into (a) *global explanations* and (b) *local explanations*. In global explanations, the emphasis is on providing a detailed overview of the model's decision-making processes (at macro-level). In essence, this approach delivers a holistic understanding of the model's behavior and how it operates across different inputs and conditions. In turn, it enhances the interpretability of the model, offering insights into its underlying operational structure and the factors that influence its overall performance during the decision-making processes [193]. Generalized Additive Model (GAM) are

 doi:10.20944/preprints202501.1423.v1

among the XAI methodologies that provide insights into a model's decision-making process at a global level [194]. GAM is a statistical modeling method designed to capture and analyze non-linear relationships between dependent and independent variables utilizing smooth functions to model the effects of each predictor [195]. For instance, the research shown in [196] utilized the GAM method to examine the relationships between kinematic variables of vehicles, such as position, velocity, and acceleration, during overtaking maneuvers. In contrast, local explanations aim to elucidate the rationale underlying specific predictions made by the model for individual instances. It is particularly valuable in situations where understanding individual predictions is important, such as analyzing specific driving scenarios in AVs. Therefore, this approach fosters trust in high-stakes autonomous systems, ensuring safety and accountability [162,193]. Grad-CAM or Gradient-weighted Class Activation Mapping is one of the prominent XAI techniques designed to interpret the decision-making process of AI models at a local level. It is often adopted to visualize and elucidate localized decisions made by CNN-based models, particularly in image recognition and classification tasks [197]. For instance, [199] adopted the Grad-CAM technique to analyze DL detection models by generating heatmaps that visually explain the road semantic segmentation outputs, thereby providing a comprehensive understanding of the relevance of their outcomes. Nevertheless, Grad-CAM may generate heatmaps that highlight regions unrelated to the detected objects in detection tasks, as its approach prioritizes feature importance without accounting for spatial sensitivity [198].

Implementation level refers to the stage at which interpretability and explainability are incorporated into AI models, focusing on when and how these aspects are integrated into the design and implementation of these models. This concept can be subdivided into (a) *ante-hoc explanations* and (b) *post-hoc explanations*. Ante-hoc explanation, also known as intrinsic explanation or pre-hoc explanation, refers to the interpretability mechanism that is inherently integrated into the design of the model during its development phase. Such explanations are designed to embed transparency and understandability into the model's decision-making processes from the outset, ensuring that its operation remains explainable and transparent from the initial stage [193,199]. Bayesian Rule Lists (BRL) represent a prominent example of an ante-hoc explanations method. It leverages Bayesian principles to achieve an optimal balance between simplicity and predictive performance. BRL operates by composing probabilistic models that derive decision rules (IF-THEN rules) based on observed data, with a focus on selecting rules that jointly maximize the posterior probability of class labels. Therefore, BRL ensures that the resulting rule lists remain explainable and grounded in a robust statistical framework [183,200]. **Figure 11** below depicts an example of how BRL can be used to explain the pedestrian crossing detection. In this instance, the model derives IF-THEN rule lists based on the input features, such as, vehicle speed, distance to pedestrians, weather conditions, and road type, to inform the decision-making process, determining whether the vehicle must stop, decelerate, or proceed with caution when detecting a potential pedestrian crossing scenario [11]. Contrarily, post-hoc explanations are applied after AI models, such as DNN or ensemble methods, have been trained. It aims to provide insights into the decision-making processes by analyzing how input features are translated into output decisions in opaque black-box models. Post-hoc explanation is critical for applications requiring model transparency, trust, and accountability, specifically when the model's complexity hinders direct interpretation [199]. Local Interpretable Model-Agnostic Explanations (LIME) is a well-known post-hoc explanation technique that approximates the decision-making processes of black-box models by constructing explainable and simplified models within the local vicinity of a specific prediction, thereby allowing stakeholders to gain insight into the reasons behind a model's decision for a particular input. For example, [201] demonstrated a trust-aware approach for selecting AVs to participate in model training, aiming to ensure system performance and reliability. They utilized the LIME method to calculate the trust values and highlight key features that influenced the selection of each AV during the model training process.

 doi:10.20944/preprints202501.1423.v1

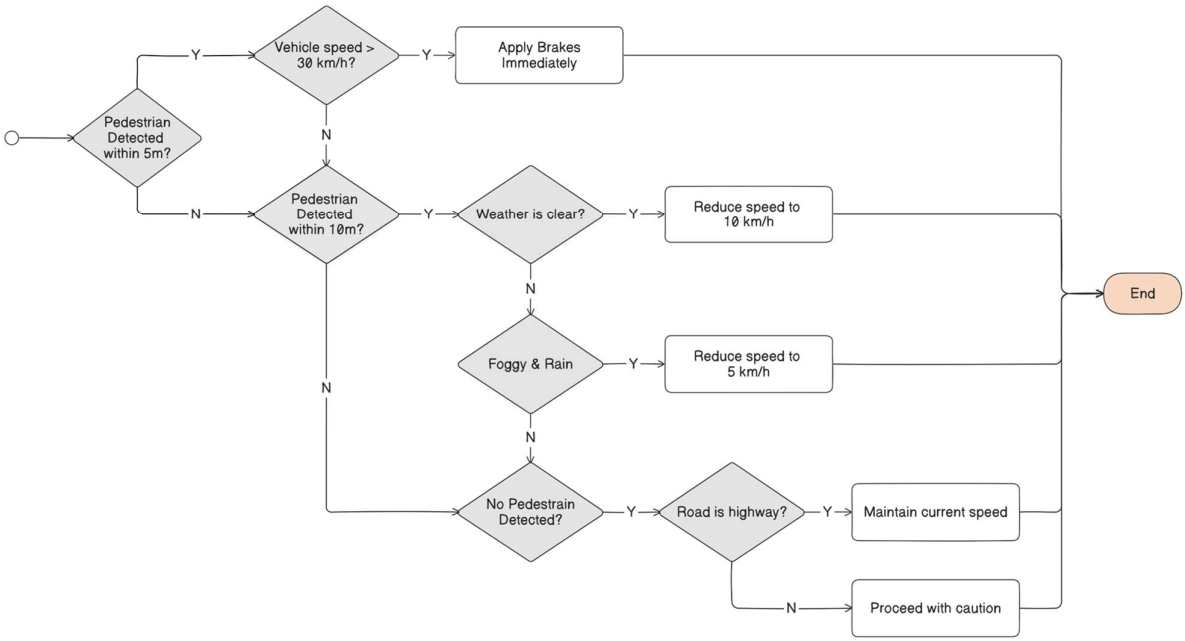


Figure 11. A graphical representation of the Bayesian Rule Lists (BRL) technique in elucidating the decision-making process for pedestrian crossing detection. The BRL rules shown in the illustration are derived in a preliminary manner based on our previous experimental analyses and discussions shown in [11]. **Rule 1:** If the pedestrian is detected within 5 m and the vehicle speed is greater than 30 km/h, then apply brakes immediately. **Rule 2:** Else if the pedestrian is detected within 10 m and the weather is clear, then reduce speed to 10 km/h. **Rule 3:** Else if the pedestrian is detected within 10 m and the weather is foggy or rain, then reduce speed to 5 km/h. **Rule 4:** Else if no pedestrian is detected and the road is highway, then maintain current speed. **Rule 5:** Else, proceed with caution.

Model dependency, as the name implies, pertains to the extent to which an explanation method is designed for a particular type of ML or DL model, or whether it possesses the versatility to be adopted across various model architectures. This concept can be subdivided into: (a) *model-agnostic* technique and (b) *model-specific* technique. Model-agnostic techniques are designed to provide interpretability independent of the underlying architecture of AI models. Model-agnostic methods are extensively utilized owing to their remarkable flexibility and adaptability, which enable them to interpret diverse models and use cases. These methods often provide post-hoc explanations and operate by examining the inputs and outputs of an AI model without requiring access to its internal parameters or structures [193,202]. Shapley Additive Explanations (SHAP) serves as a prominent example of model-agnostic explanations method. SHAP provides valuable insights into the contribution of individual input features to the output of an AI model. Moreover, it facilitates detailed and granular explanations that can either focus on specific individual predictions (local explanations) or provide an overall summary of feature importance across multiple predictions (global explanations) [203]. For instance, [204] proposed WhONet, a wheel odometry neural network that provides continuous positioning information using GNSS data with wheel encoders measurements from the vehicle. The SHAP method was adopted to interpret the predictions of vehicle positioning, thereby enhancing its reliability and ensuring greater transparency and accountability. Contrarily, model-specific techniques are designed to the unique characteristics and architecture of a specific ML or DL model. These methods leverage the intrinsic properties or mathematical properties of the model to provide detailed explanations of its decision-making processes. In other words, model-specific explanation methods require modifications to the explanation framework when applied to different models [199]. Saliency maps exemplify a model-specific interpretability technique that provides pixel-level insights into the significance of input features. This method leverages gradient-

 doi:10.20944/preprints202501.1423.v1

based information to identify and highlight the regions of an input (image) that most significantly influence the decision-making processes of an AI model by assigning a salience score to each pixel or region [205]. In other words, a saliency map represents a heatmap that highlights the most visually prominent objects or regions within a given scene. It is imperative to learn that certain studies consider that saliency maps can be generalized to operate in a model-agnostic manner by altering their computation to the model's input-output behavior rather than its internal gradients [206–208]. An illustrative application of saliency maps can be found in [209], where the authors proposed a saliency-based object detection algorithm to detect unknown obstacles in autonomous driving environments. This approach integrates the saliency map method into the detection algorithm to amplify image features, thereby emphasizing both known and unknown objects in the environment.

Table 5 and 6 below provide a detailed overview of various interpretation techniques that are commonly employed in XAI to improve the interpretability and explainability of AI models. **Table 5** categorizes these techniques based on their interpretability level (e.g., local or global), their classification within XAI (e.g., model-agnostic, model-specific, ante-hoc, and post-hoc), and the types of data they are designed to support. **Table 6** presents a comparative analysis, outlining the strengths and limitations of each interpretation technique. By consolidating this information, the tables offer valuable guidance for researchers and practitioners in identifying the most suitable techniques for specific applications. For a more in-depth exploration of additional interpretation methods (out of scope in this manuscript), readers are encouraged to refer to [167,171,179,183,184,193,194,199,210–216].

Table 5. An overview of interpretation techniques for XAI. These techniques are categorized based on their interpretability level (e.g., local or global), their explainability classification (e.g., ante-hoc, post-hoc, model-agnostic, and model-specific), and the types of input data (e.g., unstructured data – textual data, structured data – tabular, and image) that each technique can handle. The acronyms from top to bottom at the first column are: BRL – Bayesian Rule Lists; GAM – Generalized Additive Model; LIME – Local Interpretable Model-Agnostic Explanations; SHAP – Shapley Additive Explanation; Grad-CAM – Gradient-weighted Class Activation Mapping; DeepLIFT – Deep Learning Important Features; PDP – Partial Dependence Plot. This table has been adapted and revised based on [167,171,183,184,193,194,199,210–214,216].

Techniques	**Explanation Level**		**Implementation Level**		**Model Dependency**		**Data Type**		
	Global	Local	Ante-hoc	Post-hoc	Agnostic	Specific	Tabular	Image	Textual
Decision Tree	●	●	●	-	●	-	●	-	-
Linear Model	●	-	●	-	●	-	●	-	-
BRL	●	-	●	-	-	●	●	-	-
GAM	●	-	●	-	-	●	●	-	-
LIME	-	●	-	●	●	-	●	●	●
SHAP	●	●	-	●	●	-	●	●	●
Saliency Maps *	-	●	-	●	●	●	-	●	-
Grad-CAM	-	●	-	●	●	-	-	●	-
Anchors	-	●	-	●	●	-	●	●	●
DeepLIFT	●	●	-	●	●	-	-	●	●
Counterfactuals	-	●	-	●	●	-	●	●	●
Sensitivity Analysis *	●	-	-	●	●	●	●	-	-
Distillation	●	-	-	●	-	●	●	●	●
PDP	●	●	-	-	●	-	●	-	-
Feature Importance	●	●	-	●	●	-	●	●	●

* Saliency maps and sensitivity analysis can be adapted to function in a model-agnostic manner by modifying their computation to focus on the input-output relationships of a model [206–208,214].

 doi:10.20944/preprints202501.1423.v1

Table 6. A comparative analysis of interpretation techniques, highlighting their respective strengths and limitations. This table has been revised and adapted based on [167,171,183,184,193,199–201,210–216]. The acronyms from top to bottom (first column) are BRL – Bayesian Rule Lists; GAM – Generalized Additive Model; LIME – Local Interpretable Model-Agnostic Explanations; SHAP – Shapley Additive Explanations; Grad-CAM – Gradient-weighted Class Activation Mappings; DeepLIFT – Deep Learning Important Features; PDP – Partial Dependence Plot.

Techniques	Strengths	Limitations
Decision Tree	• Easy to understand. • Robust to outliers and missing values. • High interpretability. • Able to handle non-linear relationships.	• Lack of stability, where small changes in training data can result in significant variations. • Prone to overfitting. • Non-smooth decision boundaries. • Not applicable to linear relationships.
Linear Model	• Simple and easy to implement. • Computationally inexpensive. • Generalize well to new datasets with linear relationships. • Transparent, no hidden layers or complex transformations.	• Not applicable to non-linear relationships. • Oversimplified explanations may not be sufficient for safety-critical applications. • Coefficients of linear models become unstable and unreliable when input features are highly correlated. • Sensitive to outliers.
BRL	• The IF-THEN rules are easy to interpret. • Incorporation of prior knowledge, which can guide the learning process and improve model performance. • Automatic feature selection. • Can handle noisy and incomplete data by modeling uncertainty.	• High computational cost. • Difficult to model complex and high-dimensional environments. • Sensitive to noisy or incomplete data. • Not feasible for large-scale systems due to scalability issues.
GAM	• Flexible – can handle linear and non-linear relationships in data. • No black-box nature. • Provides clear and interpretable relationships between input features and predicted output. • Can include regularization techniques to control model complexity.	• Computationally intensive in large datasets or high-dimensional data. • Sensitive to smoothing parameters. • Require large sample sizes to capture non-linear patterns effectively [217]. • Risk of overfitting in highly complex data.
LIME	• Computationally efficient. • Simple and intuitive for local interpretation. • Flexible, which can be applied to any ML models. • Works well on tabular, images, and text data.	• Lacks precision in capturing global feature importance. • Sensitive to perturbations and may require hyperparameter tuning [218]. • Sensitive to small changes in data or the neighborhood around the instance [219]. • Limited to local context.
SHAP	• Versatile – can be applied to various ML models [220].	• High computational cost.

 doi:10.20944/preprints202501.1423.v1

	• Provides more accurate explanations than LIME. • Fair attribution to prevent unbiased explanations. • Can handle simple and complex models.	• Can be manipulated by adversarial attacks [221]. • May require approximations in large, complex DNN that can reduce accuracy. • Assume feature independence.
Saliency Maps	• Intuitive visualization. • Can be applied during model inference. • Effective in explaining decisions of image-based models, such as CNN. • Supports model debugging.	• Limited to gradient-based models • Sensitive to noise. • Lack of global interpretability. • Requires backpropagation, which can be computationally expensive. • Can be manipulated by adversarial attacks [222].
Grad-CAM	• Intuitive visual explanations. • Localized insights. • Robust to adversarial perturbations in image classification tasks. • Supports model debugging by highlighting which areas of the input are important for predictions.	• Lack ability to highlight fine-grained details. • Computationally expensive to calculate gradients in deep models. • Does not effectively localize objects within an image when multiple instances of the same class are present [223].
Anchors	• Less computation than SHAP. • Better generalizability than LIME [224]. • Can be applied to any ML models regardless of its architecture. • High fidelity.	• Require tuning to provide optimal explanations [225]. • Requires discretization, highly configurable, and impactful setup. • Computationally intensive.
DeepLIFT	• Compatible with DNN. • Efficient explanation generation. • Captures complex interactions between features. • Scalable. • Local and global interpretability.	• Sensitive to initialization. • Depends on a reference point or baseline, which might not always be appropriate in certain contexts. • Produce inconsistent results due to redefining gradients. • Struggle to offer global explanations for more complex and ensemble models.
Counterfactuals	• User centric – provides intuitive explanations with "what-if" scenarios. • Does not require access to the data or the model. • Easy to implement. • Provides actionable insights.	• High computational cost in high-dimensional models. • Ambiguity in interpretation and may require expert judgement in specific contexts. • Potential risk of neglecting complex relationships in data. • Inability to capture all aspects of model behavior, limiting the comprehensiveness of the explanation.
Sensitivity Analysis	• Provides intuitive explanations. • Provides unique solution, training free process, and fast computation [226]. • Identifies weak and prominent features.	• Limited to global insights. • Require explicit modeling of complex feature interactions • Computationally expensive for complex models due to multiple

 doi:10.20944/preprints202501.1423.v1

	• Applicable to various model types without requiring access to internal parameters.	evaluations for each input variation. • Generates noisy explanation maps
Distillation	• Simplifies complex models. • Can be applied across various ML models. • Does not require the creation of additional rules or decision pathways. • Maintains model's performance while ensuring interpretability.	• Dependence on the teacher model (complex model). • Potential loss of fine-grained details during compression. • Sensitive to hyperparameters. • Increase computational cost.
PDP	• Easy to implement. • Provides clear and causal interpretation. • Offers intuitive visualization. • Delivers global insights into the overall impact of individual features on predictions.	• Assumes no correlation between features. • High computational cost in large datasets. • Restricted to marginal effects, showing the influence of a maximum of three features at once. • Potential to overlook heterogeneous effects.
Feature Importance	• Provides clear and intuitive explanations. • Identifies critical factors influencing decision-making [227]. • Aids in model debugging by detecting potential biases, errors, or overfitting through feature analysis. • Offers flexibility as a model agnostic approach.	• Overlook complex feature interactions may be overlooked when decision-making processes are overly simplified. • Feature importance values are context dependent and may vary significantly across different data distributions or conditions. • High computational cost for large models. • Over-reliance on the assumption of feature independence, not suitable in scenarios where features are correlated [228,229].

3.2. Roles of XAI in Autonomous Vehicles and its Challenges

AVs are inherently complex systems, incorporating advanced and intricate AI algorithms to perceive, navigate, and make real-time decisions in dynamic, often unpredictable environments. These decisions necessitate careful consideration of numerous factors, including prevailing traffic conditions, potential road hazards, and interactions with various road users – pedestrians, cyclists, and other vehicles. However, the inherent opacity of sophisticated ML and DL models, often described as the black-box nature of AI, poses significant challenges in translating complex decision-making processes into transparent and understandable explanations, particularly in contexts where trustworthiness, safety, reliability, and accountability are imperative. For example, the rationale behind the decisions to apply brakes or swerve to avoid obstacles during autonomous driving might remain obscure to human stakeholders and may undermine the confidence in its reliability and ethical alignment. As a result, the integration of XAI holds paramount importance in addressing these challenges, as it directly impacts critical factors that are essential for the successful deployment, operation, and societal acceptance of these technologies [170].

XAI serves as a critical bridge between advanced AI-driven technologies and human understanding, providing explainable insights into the underlying decision-making processes of AI-driven systems, especially in safety-critical domains such as AVs. One of the primary roles of XAI is

 doi:10.20944/preprints202501.1423.v1

to improve transparency, which is a quintessence quality that enables human stakeholders to understand and evaluate the rationale behind the decisions made by autonomous driving systems. It demystifies the black-box nature of intricate AI algorithms and elucidates how inputs, such as sensor data, predetermined rules, and environmental conditions influence the decisions of acceleration, braking (deceleration), or navigating through complex traffic scenarios. These explanations are often presented using natural language depictions or visualizations, making the decision-making processes of autonomous driving systems more accessible and easier to interpret for diverse audiences [230]. For instance, an XAI-driven multi-sensor perception system of an AV can interpret and elucidate the relative contributions of each sensor in detecting obstacles, such as pedestrians, vehicles, or cyclists, while also providing the underlying rationale for specific decisions such as the decision to decelerate in response to detected hazards. In addition, the system may also integrate visual representations to demonstrate how sensor inputs shaped its decisions, thereby assuring end-users that the vehicle's decisions and actions are made based on robust and explainable interpretations of its environment. In instances where errors occur, XAI can assist engineers in tracing the decisions back to their originating data sources, which aids in diagnosing issues and improving detection precision; and ultimately contributes to the improved transparency and interpretability of the multi-sensor perception system [215,231–233].

However, XAI-driven systems still encounter various technical challenges that complicate their implementation and practical usability. Among these, one of the most prominent challenges is the inherent complexity of DL models, which serve as the backbone of many autonomous systems. DL models, especially DNNs, are integral to processing vast amounts of high-dimensional data and making real-time decisions. Nonetheless, their intricate architectures and reliance on sophisticated mathematical computations to achieve optimal performance in driving tasks, such as obstacle avoidance, path planning, and object detection, make it difficult to trace or elucidate the rationale behind a specific output. For example, providing an explanation for why an AV selects a particular route or reacts to hazards in a specific manner in real-time often requires advanced interpretability techniques, which are essential to achieve the level of explainability demanded in safety-critical systems for trust and accountability. Thus, achieving an optimal balance between interpretability and model performances remains an ongoing challenge in the development of XAI-driven systems [184]. Other technical challenges involve the need to explain real-time decisions in time-sensitive and safety-critical situations without introducing significant delays that could compromise the system's performance. For instance, in multi-sensor systems, establishing a unified framework to incorporate multimodal data sources and elucidate the contribution of each sensor modality in real-time is a significant challenge as these systems scale in size to address various driving conditions. There is also the potential computational overhead associated with generating interpretable explanations without affecting real-time performances. Moreover, the challenge of establishing a universal explanation technique that applies to diverse and dynamic environments remains significant. This includes the difficulty of explaining decisions made in edge cases or unprecedented conditions, as well as the need to generalize explanations across different driving scenarios, operational contexts, stakeholder groups, and modes of transport (on-road versus off-road) [168,170].

Transparency, in turn, supports trustworthiness, which is a critical factor in promoting the widespread acceptance and successful adoption of AI-driven systems across various domains. In the early stages of technological advancement, machines and algorithms were often viewed as epitomes of trustworthiness and reliability due to their predictable, as their operations and actions were limited to executing predefined tasks that are explicitly programmed, leaving minimal scope for ambiguity or error in their decision-making processes. In recent years, the emergence of ML and DL algorithms has marked a significant paradigm shift, facilitating the creation of systems capable of autonomous reasoning and decision-making. However, this evolution has also introduced an element of unpredictability and opacity into the behavior of AI-driven systems, which in turn undermines the implicit trust due to the underlying complex and opaque reasoning behind their decisions [234]. From end-users' perspective, the concept of trustworthiness in these systems extends beyond their

 doi:10.20944/preprints202501.1423.v1

technical capabilities. It operates as a socio-psychological construct that impacts how individuals, communities, and societies perceive, interact with, and ultimately accept emerging technologies, specifically in autonomous driving systems [235]. One primary factor that affects trustworthiness from a socio-psychological perspective is the fear of the unknown, which stems from the inherent complexity and unpredictability of these technologies. This concern is especially significant in safety-critical applications, where system failure or malfunctions can result in severe and far-reaching consequences. Besides, the lack of clear accountability in autonomous systems intensifies the fear of the unknown, creating significant uncertainty regarding responsibility in the event of system failures or accidents. Thus, the ambiguity surrounding liability and responsibility amplifies public apprehension and undermines trust in AI-driven applications [236]. Other socio-psychological factors influencing trustworthiness of AI systems include perceived behavioral control, which relates to the user's capabilities to control or intervene the system when necessitated, privacy concerns, and perceived usefulness, which refers to the belief that the system will effectively achieve its intended purposes [235]. Thus, it is imperative to highlight transparency and explainability as the foundational elements of trustworthy AI [237].

From a regulatory perspective, the capability to provide explainable insights into AI systems has emerged as an imperative requirement across multiple jurisdictions. As AVs and other AI-driven systems become increasingly integrated into various aspects of society, regulatory authorities have emphasized the critical importance of ensuring transparency and interpretability in their decision-making processes. Thus, the integration of XAI into such applications is important to complying with regulatory mandates and industry standards, as it provides critical mechanisms for comprehending, justifying, and validating the decisions and actions made by AI-driven systems. In addition, it plays an imperative role in supporting transparent investigations and aiding in the determination of liability in the event of an incident [184,238]. In April 2019, the High-Level Expert Group on AI (AI HLEG), appointed by the European Commission (EC), presented a human-centric approach for AI development, which outlines seven ethical guidelines aimed at supporting the development of AI systems that can be considered as trustworthy. **Table 7** below outlines the seven ethical guidelines that AI systems must adhere to be deemed trustworthy [234,239,240]. Moreover, XAI is essential in addressing biases within autonomous systems, specifically in instances where such biases stem from unrepresentative training data or flawed algorithmic designs. By enhancing the transparency of the AI decision-making processes, XAI enables the identification and analysis of potential sources of bias that can lead to inequitable or unfair outcomes. This capability ensures that AI-driven systems operate in a fair and unprejudiced manner, thereby preventing the perpetuation of discriminatory practices and promoting unbiased decision-making [162,183]. Nonetheless, one of the ethical challenges of XAI is that it can be challenging to identify the appropriate level of explanation required for different scenarios. Therefore, it is essential to tailor explanations that suit the unique needs and expectations of different uses cases, thereby addressing the distinct requirements of various stakeholders [241]. Furthermore, the ethical challenges associated with data security and data privacy in XAI are significant and multifaceted. It requires an optimal balance between openness and confidentiality, certifying that sensitive data is not compromised or exposed to vulnerabilities, while simultaneously ensuring that the explanations provided are clear, interpretable, and meaningful [234,235].

Table 7. An overview of the seven essential criteria outlined in the established ethical guidelines by the High-Level Expert Group on AI (AI HLEG) that AI systems must follow to be deemed as trustworthy. This table has been revised and adapted based on [234,239,240].

Criteria	Explanations
Human Agency and Oversight	AI systems should enhance human decision-making and support fundamental rights while ensuring adequate oversight, rather than restricting or misleading human autonomy. This can be achieved through human-in-the-

 doi:10.20944/preprints202501.1423.v1

	loop, human-on-the-loop, and human-in-command approaches.
Technical Robustness and Safety	AI systems must be resilient, secure, and safe, with contingency plans in place to address system failures or malfunctions. They must also be accurate, reliable, and reproducible to minimize and prevent unintentional harm.
Privacy and Data Governance	In addition to safeguarding privacy and data protection, effective data governance mechanisms must be established, ensuring data quality, integrity, and authorized access. End-users should also maintain full control over their personal information, ensuring that such data is not used in ways that could be detrimental or harmful to their interests.
Transparency	Data, systems, and AI business models must be transparent, with traceability mechanisms ensuring accountability. Moreover, AI systems and their decisions should be explained in a way that is tailored to the relevant stakeholders, and it is essential that users are aware that they are interacting with AI and are informed of its capabilities and limitations.
Diversity, Non-Discrimination, and Fairness	Unfair bias must be eliminated to prevent negative outcomes such as the marginalization of vulnerable groups and the reinforcement of prejudice. AI systems should be accessible to all, regardless of disability, and involve relevant stakeholders throughout their lifecycle to promote inclusivity.
Societal and Environmental Well-Being	AI systems must be designed to benefit all humanity, including future generations, while prioritizing sustainability and environmental responsibility. Additionally, their impact on the environment, other living being, and society must be thoroughly evaluated and considered.
Accountability	Mechanisms must be established to ensure accountability for AI systems and their outcomes. Auditability, which allows for the evaluation of algorithms, data, and design processes, is essential, particularly in critical applications. Besides, accessible avenues for compensation should be provided.

4. Conclusions and Future Research Recommendations

In this manuscript, we investigated and explored the intersection of multi-sensor fusion and XAI, aiming on addressing the challenges associated with developing interpretable, trustworthy, and accurate AV systems. We began the survey by introducing the various applications of AVs in both on-road and off-road environments, and an overview of the commonly employed sensors integral to developing multi-sensor perception systems, which support critical functionalities, including object detection, obstacle avoidance, and localization and mapping. Subsequently, we presented a comprehensive overview of the various multi-sensor fusion strategies, highlighting their respective strengths and limitations. It gave valuable insights into the various fusion approaches from three primary aspects: (a) **when** should the sensor fusion occur, (b) **where** should the sensor fusion occur; and (c) **what** should the fusion do. Ultimately, selecting the most suitable approaches depends on the specific use cases, requirements, and available resources on the AVs. Additionally, we reviewed some of the cutting-edge multi-sensor fusion techniques and algorithms – traditional and advanced fusion algorithms, discussing their respective applications, strengths, and weaknesses. We also emphasized

 doi:10.20944/preprints202501.1423.v1

the challenges involved in the deployment of reliable, safe, scalable, transparent, and comprehensible multi-sensor perception systems in real-world autonomous driving environments. Some of the key challenges are:

- Sensor noise, which relates to the inaccuracies, inconsistencies, or irrelevant data introduced by individual sensors due to a combination of hardware limitations, external interference, or environmental conditions.
- Heterogeneity of sensor modalities in AVs and the resulting system complexity.
- Achieving an optimal balance between accuracy and computational efficiency.
- Multi-sensor fusion systems are susceptible to malicious attacks, which pose significant risk to the integrity and reliability of their autonomous operation.
- Lack of transparency, explainability, and interpretability in black-box AI models, especially in advanced DNN algorithms.

Finally, we explored the core principles of XAI and provided a comprehensive overview of the several emerging XAI strategies and techniques that can be integrated during autonomous systems development to enhance the transparency, trustworthiness, and interpretability of these systems. We summarized the strengths and limitations of these approaches, offering valuable guidance for researchers and practitioners in identifying and selecting the most suitable strategies and methodologies for specific use cases. Moreover, we examined the significance of XAI in AI-driven systems, specifically in AVs, as well as the challenges associated with integrating XAI into real-time autonomous driving applications or other AI-driven technologies. The findings revealed that the lack of interpretability and transparency in advanced AI models, specifically in DNNs, remains a primary challenge due to the opaque, black-box nature of their model architectures and the inherent complexity of these systems. Eventually, the selection of suitable strategies and methodologies for incorporating XAI depends on the specific system requirements, computational resources, and the associated limitations, all while striving to attain an optimal balance between explainability and system performance. Moreover, several challenges comprise technical, ethical, social, and regulatory aspects, remain a main challenge that must be addressed to enable the successful deployment of XAI systems into the real-world environments while ensuring that such systems remain efficient, safe, transparent, trustworthy, and ethical.

In summary, the development of methodologies that ensure real-time explainability for stakeholders without compromising safety and accuracy is paramount in the successful deployment of AVs and other AI-driven systems. It ensures that stakeholders, including end-users, engineers, operators, and regulators, can understand the reasoning behind critical decisions while it operates in complex and dynamic environments, fostering trust and enabling timely interventions when needed. Nonetheless, it is essential to customize the explanations to meet the specific needs and expectations of different use cases, thereby addressing the diverse requirements of various stakeholders. In autonomous driving, vehicles operate in real-time and must adapt to rapidly changing situations. It is imperative to attain an optimal balance between the computational requirements necessitated for accurate real-time decision-making and the need for explainability and transparency, without introducing delays that could result in potential hazardous outcomes. Hence, it is essential to develop efficient and scalable XAI methods that provide clear, comprehensible, and real-time explanations, while maintaining operational safety and decision-making accuracy of autonomous systems. Such methods are critical for fostering trust and accountability, aiding in error diagnosis, ensuring compliance with regulatory requirements, and supporting the ethical and responsible integration and deployment of autonomous technologies into real-world environments.

Future research directions aimed at progressing the integration of XAI into real-time, high-stakes AVs or other AI-driven systems encompass a range of innovative and critical domains. Such explorations aim to address existing challenges and unlock new opportunities to enhance the safety, reliability, interpretability, transparency, and trustworthiness of these systems. A significant area of focus for future research involves the development of a unified context-aware evaluation framework for comparing and selecting interpretability techniques across multiple domains or, at a minimum,

 doi:10.20944/preprints202501.1423.v1

achieving uniformity within specialized areas. It could contribute to the development of best practices in XAI, providing valuable, contextual, and adaptive insights that are aligned with specific goals, stakeholders, and operational constraints of different domains – cross-disciplinary, human-AI collaboration [170]. Over time, this would support the development of more transparent, reliable, and user-centric AI systems [184,242,243]. Moreover, it is essential to investigate and develop novel XAI approaches that facilitate the provision of accurate and computationally efficient real-time explanations, specifically in memory-constrained, real-time industrial systems like autonomous driving and healthcare [171]. Another promising direction for future research involves integrating causal relationships into XAI, with the objective of enhancing the capability of AI systems to offer more comprehensive explanations for their decisions. This approach aims to elucidate the underlying causal factors that impact the outcomes, thereby enabling a transparent understanding of the cause-and-effect dynamics involved in the decision-making process [171,216,244,245].

Besides, it is important to investigate and refine cutting-edge multi-sensor fusion algorithms capable of processing and interpreting large-scale sensor data in real time. Such advancements are vital to ensuring the accuracy and reliability of autonomous systems in dynamic environments, while simultaneously providing clear and interpretable explanations of the underlying decision-making process. From an ethical and regulatory perspective, future research should prioritize the development of methodologies aimed at incorporating fairness, non-discrimination, and privacy protections into AI systems. Simultaneously, it is vital to ensure that these systems comply with emerging ethical and regulatory standards, thereby fostering trust and accountability within AI technology [171,246]. Other future research avenues may involve incorporating large language models (LLMs) to aid in the generation of clear, contextually relevant, and user-friendly explanations for various stakeholders, including passengers, regulators, and legal professionals [247,248]. Moreover, investigating the different methodologies for preventing adversarial attacks is vital in ensuring the security and integrity of AI systems, specifically in safety-critical applications [240,249,250]. Finally, improving the knowledge and skills of practitioners and researchers in XAI through continuous education and training will significantly contribute to the advancement of interpretability research and its practical applications. It is also important to develop accessible and effective educational frameworks aimed at fostering public understanding of AI systems, their capabilities and limitations, as well as their decision-making processes [184,235]. We hope that these research avenues will facilitate the development of AI models that are reliable, trustworthy, interpretable, and safe, thereby advancing the field of XAI and enhancing transparency and interpretability in AVs.

Author Contributions: Conceptualization, D.J.Y.; methodology, D.J.Y.; software, D.J.Y.; validation, D.J.Y., K.P.; formal analysis, D.J.Y.; investigation, D.J.Y.; resources, D.J.Y.; data curation, D.J.Y.; writing—original draft preparation, D.J.Y.; writing—review and editing, D.J.Y., K.P., J.W.; visualization, D.J.Y.; supervision, K.P., J.W.; project administration, K.P., J.W.; funding acquisition, K.P., J.W.. All authors have read and agreed to the published version of the manuscript.

Funding: The research reported in this article was supported, in part, by Taighde Éireann – Research Ireland under Grant number 13/RC/2094_P2 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero – the Research Ireland Centre for Software (www.lero.ie).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: This article and the research behind it would not have been possible without the support of the IMaR team in the Munster Technological University, Co. Kerry.

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

 doi:10.20944/preprints202501.1423.v1

Abbreviations

The following abbreviations are used in this manuscript:

3D	Three Dimensional
AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on AI
AV	Autonomous Vehicles
BEV	Bird's-Eye View
BRL	Bayesian Rule Lists
CNN	Convolutional Neural Networks
DeepLIFT	Deep Learning Important Features
DL	Deep Learning
DNN	Deep Neural Network
DST	Dempster-Shafer Theory
EC	European Commission
EM	Expectation-Maximization
Faster R-CNN	Faster Region-Convolutional Neural Network
GAM	Generalized Additive Model
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
HD	High-Definition
HLF	High-Level Fusion
IMU	Inertial Measurement Unit
IoU	Intersection over Union
KF	Kalman Filter
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-Agnostic Explanations
LLF	Low-Level Fusion
LLM	Large Language Model
ML	Machine Learning
MLF	Mid-Level Fusion
MMLF	Mult-modal Multi-class Late Fusion
NMS	Non-Maximum Suppression
PDP	Partial Dependency Plots
PF	Particle Filter
RBM	Restricted Boltzmann Machine
RL	Reinforcement Learning
RMG	Rail Mounted Gantry
RNN	Recurrent Neural Networks
RPN	Region Proposal Network
SAE	Society of Automation Engineers
SCFT	Spatio-Contextual Fusion Transformer
SHAP	Shapley Additive Explanations
SPA	Soft Polar Association
TPU	Tensor Processing Unit
UKF	Unscented Kalman Filter
XAI	Explainable Artificial Intelligence

References

1. Vemoori, V. Towards Safe and Equitable Autonomous Mobility: A Multi-Layered Framework Integrating Advanced Safety Protocols, Data-Informed Road Infrastructure, and Explainable AI for Transparent Decision-Making in Self-Driving Vehicles. *Human-Computer Interaction Persp.* **2022**, *2*, 10–41.
2. Olayode, I.O.; Du, B.; Severino, A.; Campisi, T.; Alex, F.J. Systematic Literature Review on the Applications, Impacts, and Public Perceptions of Autonomous Vehicles in Road Transportation System. *J. Traffic Transp. Eng. Engl. Ed.* **2023**, *10*, 1037–1060, doi:10.1016/j.jtte.2023.07.006.
3. Velasco-Hernandez, G.; Yeong, D.J.; Barry, J.; Walsh, J. Autonomous Driving Architectures, Perception and Data Fusion: A Review. In Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP); September 2020; pp. 315–321.
4. Rondelli, V.; Franceschetti, B.; Mengoli, D. A Review of Current and Historical Research Contributions to the Development of Ground Autonomous Vehicles for Agriculture. *Sustainability* **2022**, *14*, 9221, doi:10.3390/su14159221.
5. Chen, L.; Li, Y.; Silamu, W.; Li, Q.; Ge, S.; Wang, F.-Y. Smart Mining With Autonomous Driving in Industry 5.0: Architectures, Platforms, Operating Systems, Foundation Models, and Applications. *IEEE Trans. Intell. Veh.* **2024**, *9*, 4383–4393, doi:10.1109/TIV.2024.3365997.
6. Molina, A.A.; Huang, Y.; Jiang, Y. A Review of Unmanned Aerial Vehicle Applications in Construction Management: 2016–2021. *Standards* **2023**, *3*, 95–109, doi:10.3390/standards3020009.
7. Olapoju, O.M. Autonomous Ships, Port Operations, and the Challenges of African Ports. *Marit. Technol. Res.* **2023**, *5*, 260194–260194, doi:10.33175/mtr.2023.260194.
8. Naeem, D.; Gheith, M.; Eltawil, A. A Comprehensive Review and Directions for Future Research on the Integrated Scheduling of Quay Cranes and Automated Guided Vehicles and Yard Cranes in Automated Container Terminals. *Comput. Ind. Eng.* **2023**, *179*, 109149, doi:10.1016/j.cie.2023.109149.
9. Negash, N.M.; Yang, J. Driver Behavior Modeling Toward Autonomous Vehicles: Comprehensive Review. *IEEE Access* **2023**, *11*, 22788–22821, doi:10.1109/ACCESS.2023.3249144.
10. Zhao, J.; Zhao, W.; Deng, B.; Wang, Z.; Zhang, F.; Zheng, W.; Cao, W.; Nan, J.; Lian, Y.; Burke, A.F. Autonomous Driving System: A Comprehensive Survey. *Expert Syst. Appl.* **2024**, *242*, 122836, doi:10.1016/j.eswa.2023.122836.
11. Yeong, D.J.; Velasco-Hernandez, G.; Barry, J.; Walsh, J. Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review. *Sensors* **2021**, *21*, 2140, doi:10.3390/s21062140.
12. Zhang, Y.; Carballo, A.; Yang, H.; Takeda, K. Perception and Sensing for Autonomous Vehicles under Adverse Weather Conditions: A Survey. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 146–177, doi:10.1016/j.isprsjprs.2022.12.021.
13. Waymo - Self-Driving Cars - Autonomous Vehicles - Ride-Hail. Available online: https://waymo.com/ (accessed on 25 October 2024).
14. Einride - Intelligent Movement. Available online: https://einride.tech/ (accessed on 21 October 2024).
15. Automated RMG (ARMG) System | Konecranes. Available online: https://www.konecranes.com/port-equipment-services/container-handling-equipment/automated-rmg-armg-system (accessed on 21 October 2024).
16. Autonomous Tractor | John Deere US. Available online: https://www.deere.com/en/autonomous/ (accessed on 21 October 2024).
17. Autonomous Pallet Loader - Stratom. Available online: https://www.stratom.com/apl/ (accessed on 21 October 2024).
18. *FLIR Thermal Imaging Enables Autonomous Inspections of Mining Vehicles*; 2020;
19. Ludlow, E. Waymo Sets Its Sights on 'Premium' Robotaxi Passengers. *Bloomberg.com* 2024.
20. SAE J3016 Automated-Driving Graphic. Available online: https://www.sae.org/site/news/2019/01/sae-updates-j3016-automated-driving-graphic (accessed on 22 October 2024).
21. J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - SAE International. Available online: https://www.sae.org/standards/content/j3016_202104/ (accessed on 22 October 2024).

22. Infographic: Cars Increasingly Ready for Autonomous Driving. Available online: https://www.statista.com/chart/25754/newly-registered-cars-by-autonomous-driving-level (accessed on 22 October 2024).
23. Ackerman, E. What Full Autonomy Means for the Waymo Driver - IEEE Spectrum. Available online: https://spectrum.ieee.org/full-autonomy-waymo-driver (accessed on 22 October 2024).
24. SAE Levels of Driving Automation™ Refined for Clarity and International Audience. Available online: https://www.sae.org/site/blog/sae-j3016-update (accessed on 22 October 2024).
25. Press, R. SAE Levels of Automation in Cars Simply Explained (+Image). Available online: https://www.rambus.com/blogs/driving-automation-levels/ (accessed on 28 October 2024).
26. Raciti, M.; Bella, G. A Threat Model for Soft Privacy on Smart Cars 2023.
27. Senel, N.; Kefferpütz, K.; Doycheva, K.; Elger, G. Multi-Sensor Data Fusion for Real-Time Multi-Object Tracking. *Processes* **2023**, *11*, 501, doi:10.3390/pr11020501.
28. Xiang, C.; Feng, C.; Xie, X.; Shi, B.; Lu, H.; Lv, Y.; Yang, M.; Niu, Z. Multi-Sensor Fusion and Cooperative Perception for Autonomous Driving: A Review. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 36–58, doi:10.1109/MITS.2023.3283864.
29. Dong, J.; Chen, S.; Miralinaghi, M.; Chen, T.; Li, P.; Labi, S. Why Did the AI Make That Decision? Towards an Explainable Artificial Intelligence (XAI) for Autonomous Driving Systems. *Transp. Res. Part C Emerg. Technol.* **2023**, *156*, 104358, doi:10.1016/j.trc.2023.104358.
30. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Towards Safe, Explainable, and Regulated Autonomous Driving. 2023.
31. Data Analysis: Self-Driving Car Accidents [2019-2024]. *Craft Law Firm*.
32. Muhammad, K.; Ullah, A.; Lloret, J.; Ser, J.D.; de Albuquerque, V.H.C. Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 4316–4336, doi:10.1109/TITS.2020.3032227.
33. Port & Terminal Automation | Quanergy Solutions, Inc. | LiDAR Sensors and Smart Perception Solutions. Available online: https://quanergy.com/port-terminal-automation/ (accessed on 8 November 2024).
34. Autonomous Vehicles Factsheet | Center for Sustainable Systems. Available online: https://css.umich.edu/publications/factsheets/mobility/autonomous-vehicles-factsheet (accessed on 8 November 2024).
35. Odukha, O. How Sensor Fusion for Autonomous Cars Helps Avoid Deaths on the Road. Available online: https://intellias.com/sensor-fusion-autonomous-cars-helps-avoid-deaths-road/ (accessed on 8 November 2024).
36. Introduction to Autonomous Driving Sensors [BLOG] | News | ECOTRONS. Available online: https://ecotrons.com/news/introduction-to-autonomous-driving-sensors-blog/ (accessed on 8 November 2024).
37. Grey, T. The Anatomy of an Autonomous Vehicle | Ouster. Available online: https://ouster.com/insights/blog/the-anatomy-of-an-autonomous-vehicle (accessed on 8 November 2024).
38. Dreissig, M.; Scheuble, D.; Piewak, F.; Boedecker, J. Survey on LiDAR Perception in Adverse Weather Conditions. 2023.
39. Kim, J.; Park, B.; Kim, J. Empirical Analysis of Autonomous Vehicle's LiDAR Detection Performance Degradation for Actual Road Driving in Rain and Fog. *Sensors* **2023**, *23*, 2972, doi:10.3390/s23062972.
40. Park, J.; Cho, J.; Lee, S.; Bak, S.; Kim, Y. An Automotive LiDAR Performance Test Method in Dynamic Driving Conditions. *Sensors* **2023**, *23*, 3892, doi:10.3390/s23083892.
41. Wang, Z.; Wu, Y.; Niu, Q. Multi-Sensor Fusion in Automated Driving: A Survey. *IEEE Access* **2020**, *8*, 2847–2868, doi:10.1109/ACCESS.2019.2962554.
42. Cao, Y.; Wang, N.; Xiao, C.; Yang, D.; Fang, J.; Yang, R.; Chen, Q.A.; Liu, M.; Li, B. Invisible for Both Camera and LiDAR: Security of Multi-Sensor Fusion Based Perception in Autonomous Driving Under Physical-World Attacks. 2021.
43. Hafeez, F.; Sheikh, U.U.; Alkhaldi, N.; Garni, H.Z.A.; Arfeen, Z.A.; Khalid, S.A. Insights and Strategies for an Autonomous Vehicle With a Sensor Fusion Innovation: A Fictional Outlook. *IEEE Access* **2020**, *8*, 135162–135175, doi:10.1109/ACCESS.2020.3010940.

44. Hasanujjaman, M.; Chowdhury, M.Z.; Jang, Y.M. Sensor Fusion in Autonomous Vehicle with Traffic Surveillance Camera System: Detection, Localization, and AI Networking. *Sensors* **2023**, *23*, 3335, doi:10.3390/s23063335.
45. Marti, E.; de Miguel, M.A.; Garcia, F.; Perez, J. A Review of Sensor Technologies for Perception in Automated Driving. *IEEE Intell. Transp. Syst. Mag.* **2019**, *11*, 94–108, doi:10.1109/MITS.2019.2907630.
46. Butt, F.A.; Chattha, J.N.; Ahmad, J.; Zia, M.U.; Rizwan, M.; Naqvi, I.H. On the Integration of Enabling Wireless Technologies and Sensor Fusion for Next-Generation Connected and Autonomous Vehicles. *IEEE Access* **2022**, *10*, 14643–14668, doi:10.1109/ACCESS.2022.3145972.
47. Yoon, K.; Choi, J.; Huh, K. Adaptive Decentralized Sensor Fusion for Autonomous Vehicle: Estimating the Position of Surrounding Vehicles. *IEEE Access* **2023**, *11*, 90999–91008, doi:10.1109/ACCESS.2023.3308152.
48. Thakur, A.; Mishra, S.K. An In-Depth Evaluation of Deep Learning-Enabled Adaptive Approaches for Detecting Obstacles Using Sensor-Fused Data in Autonomous Vehicles. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108550, doi:10.1016/j.engappai.2024.108550.
49. Argese, S. Sensors Selection for Obstacle Detection: Sensor Fusion and YOLOv4 for an Autonomous Surface Vehicle in Venice Lagoon. laurea, Politecnico di Torino, 2023.
50. Matos, F.; Bernardino, J.; Durães, J.; Cunha, J. A Survey on Sensor Failures in Autonomous Vehicles: Challenges and Solutions. *Sensors* **2024**, *24*, 5108, doi:10.3390/s24165108.
51. Brena, R.F.; Aguileta, A.A.; Trejo, L.A.; Molino-Minero-Re, E.; Mayora, O. Choosing the Best Sensor Fusion Method: A Machine-Learning Approach. *Sensors* **2020**, *20*, 2350, doi:10.3390/s20082350.
52. Dong, H.; Gu, W.; Zhang, X.; Xu, J.; Ai, R.; Lu, H.; Kannala, J.; Chen, X. SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); May 2024; pp. 9056–9062.
53. Ku, J.; Harakeh, A.; Waslander, S.L. In Defense of Classical Image Processing: Fast Depth Completion on the CPU. In Proceedings of the 2018 15th Conference on Computer and Robot Vision (CRV); May 2018; pp. 16–22.
54. Kim, Y.; Kim, S.; Choi, J.W.; Kum, D. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 1160–1168, doi:10.1609/aaai.v37i1.25198.
55. nuScenes Available online: https://www.nuscenes.org/ (accessed on 10 November 2024).
56. Srivastav, A.; Mandal, S. Radars for Autonomous Driving: A Review of Deep Learning Methods and Challenges. *IEEE Access* **2023**, *11*, 97147–97168, doi:10.1109/ACCESS.2023.3312382.
57. Fawole, O.A.; Rawat, D.B. Recent Advances in 3D Object Detection for Self-Driving Vehicles: A Survey. *AI* **2024**, *5*, 1255–1285, doi:10.3390/ai5030061.
58. Brabandere, B.D. Late vs Early Sensor Fusion for Autonomous Driving. *Segments.ai* 2024.
59. Shi, K.; He, S.; Shi, Z.; Chen, A.; Xiong, Z.; Chen, J.; Luo, J. Radar and Camera Fusion for Object Detection and Tracking: A Comprehensive Survey. 2024.
60. Pandharipande, A.; Cheng, C.-H.; Dauwels, J.; Gurbuz, S.Z.; Ibanez-Guzman, J.; Li, G.; Piazzoni, A.; Wang, P.; Santra, A. Sensing and Machine Learning for Automotive Perception: A Review. *IEEE Sens. J.* **2023**, *23*, 11097–11115, doi:10.1109/JSEN.2023.3262134.
61. Sural, S.; Sahu, N.; Rajkumar, R. ContextualFusion: Context-Based Multi-Sensor Fusion for 3D Object Detection in Adverse Operating Conditions. 2024.
62. Huch, S.; Sauerbeck, F.; Betz, J. DeepSTEP - Deep Learning-Based Spatio-Temporal End-To-End Perception for Autonomous Vehicles. In Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV); IEEE, June 2023; pp. 1–8.
63. Visteon | Current Sensor Data Fusion Architectures: Visteon's Approach. Available online: https://www.visteon.com/current-sensor-data-fusion-architectures-visteons-approach/ (accessed on 18 November 2024).
64. Singh, A. Vision-RADAR Fusion for Robotics BEV Detections: A Survey. 2023.
65. Jahn, L.L.F.; Park, S.; Lim, Y.; An, J.; Choi, G. Enhancing Lane Detection with a Lightweight Collaborative Late Fusion Model. *Robot. Auton. Syst.* **2024**, *175*, 104680, doi:10.1016/j.robot.2024.104680.
66. Yang, Q.; Zhao, Y.; Cheng, H. MMLF: Multi-Modal Multi-Class Late Fusion for Object Detection with Uncertainty Estimation. 2024.

67. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision Meets Robotics: The KITTI Dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237, doi:10.1177/0278364913491297.
68. Park, J.; Thota, B.K.; Somashekar, K. Sensor-Fused Nighttime System for Enhanced Pedestrian Detection in ADAS and Autonomous Vehicles. *Sensors* **2024**, *24*, 4755, doi:10.3390/s24144755.
69. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645, doi:10.1109/TPAMI.2009.167.
70. 9 Types of Sensor Fusion Algorithms. Available online: https://www.thinkautonomous.ai/blog/9-types-of-sensor-fusion-algorithms/ (accessed on 29 November 2024).
71. Hasanujjaman, M.; Chowdhury, M.Z.; Jang, Y.M. Sensor Fusion in Autonomous Vehicle with Traffic Surveillance Camera System: Detection, Localization, and AI Networking. *Sensors* **2023**, *23*, 3335, doi:10.3390/s23063335.
72. Anisha, A.M.; Abdel-Aty, M.; Abdelraouf, A.; Islam, Z.; Zheng, O. Automated Vehicle to Vehicle Conflict Analysis at Signalized Intersections by Camera and LiDAR Sensor Fusion. *Transp. Res. Rec. J. Transp. Res. Board* **2023**, *2677*, 117–132, doi:10.1177/03611981221128806.
73. Parida, B. Sensor Fusion: The Ultimate Guide to Combining Data for Enhanced Perception and Decision-Making. Available online: https://www.wevolver.com/article/sensor-fusion, https://www.wevolver.com/article/sensor-fusion (accessed on 30 November 2024).
74. Dorlecontrols. Sensor Fusion. *Medium* 2023.
75. Thakur, A.; Mishra, S.K. An In-Depth Evaluation of Deep Learning-Enabled Adaptive Approaches for Detecting Obstacles Using Sensor-Fused Data in Autonomous Vehicles. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108550, doi:10.1016/j.engappai.2024.108550.
76. Fayyad, J.; Jaradat, M.A.; Gruyer, D.; Najjaran, H. Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review. *Sensors* **2020**, *20*, 4220, doi:10.3390/s20154220.
77. Stateczny, A.; Wlodarczyk-Sielicka, M.; Burdziakowski, P. Sensors and Sensor's Fusion in Autonomous Vehicles. *Sensors* **2021**, *21*, 6586, doi:10.3390/s21196586.
78. Tian, C.; Liu, H.; Liu, Z.; Li, H.; Wang, Y. Research on Multi-Sensor Fusion SLAM Algorithm Based on Improved Gmapping. *IEEE Access* **2023**, *11*, 13690–13703, doi:10.1109/ACCESS.2023.3243633.
79. Ding, Z.; Sun, Y.; Xu, S.; Pan, Y.; Peng, Y.; Mao, Z. Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing. *Robotics* **2023**, *12*, 100, doi:10.3390/robotics12040100.
80. Park, J.; Kim, C.; Kim, S.; Jo, K. PCSCNet: Fast 3D Semantic Segmentation of LiDAR Point Cloud for Autonomous Car Using Point Convolution and Sparse Convolution Network. *Expert Syst. Appl.* **2023**, *212*, 118815, doi:10.1016/j.eswa.2022.118815.
81. Wang, Y.; Mao, Q.; Zhu, H.; Deng, J.; Zhang, Y.; Ji, J.; Li, H.; Zhang, Y. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey. *Int. J. Comput. Vis.* **2023**, *131*, 2122–2152, doi:10.1007/s11263-023-01784-z.
82. Mao, J.; Shi, S.; Wang, X.; Li, H. 3D Object Detection for Autonomous Driving: A Comprehensive Survey. *Int. J. Comput. Vis.* **2023**, *131*, 1909–1963, doi:10.1007/s11263-023-01790-1.
83. Appiah, E.O.; Mensah, S. Object Detection in Adverse Weather Condition for Autonomous Vehicles. *Multimed. Tools Appl.* **2024**, *83*, 28235–28261, doi:10.1007/s11042-023-16453-z.
84. Alaba, S.Y.; Gurbuz, A.C.; Ball, J.E. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. *World Electr. Veh. J.* **2024**, *15*, 20, doi:10.3390/wevj15010020.
85. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117, doi:10.1016/j.neunet.2014.09.003.
86. Gurney, K. *An Introduction to Neural Networks*; 1st ed.; CRC Press: London, 2018; ISBN 978-1-315-27357-0.
87. Fan, F.; Xiong, J.; Li, M.; Wang, G. On Interpretability of Artificial Neural Networks: A Survey. 2021.
88. Alaba, S.Y. GPS-IMU Sensor Fusion for Reliable Autonomous Vehicle Position Estimation. 2024.
89. Tian, K.; Radovnikovich, M.; Cheok, K. Comparing EKF, UKF, and PF Performance for Autonomous Vehicle Multi-Sensor Fusion and Tracking in Highway Scenario. In Proceedings of the 2022 IEEE International Systems Conference (SysCon); April 2022; pp. 1–6.

90. Aamir, M.; Nawi, N.M.; Wahid, F.; Mahdin, H. An Efficient Normalized Restricted Boltzmann Machine for Solving Multiclass Classification Problems. *Int. J. Adv. Comput. Sci. Appl. IJACSA* **2019**, *10*, doi:10.14569/IJACSA.2019.0100856.
91. Li, L.; Sheng, X.; Du, B.; Wang, Y.; Ran, B. A Deep Fusion Model Based on Restricted Boltzmann Machines for Traffic Accident Duration Prediction. *Eng. Appl. Artif. Intell.* **2020**, *93*, 103686, doi:10.1016/j.engappai.2020.103686.
92. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A.A.A.; Yogamani, S.; Pérez, P. Deep Reinforcement Learning for Autonomous Driving: A Survey. 2021.
93. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021.
94. Lu, D.; Xie, Q.; Wei, M.; Gao, K.; Xu, L.; Li, J. Transformers in 3D Point Clouds: A Survey. 2022.
95. Li, P.; Pei, Y.; Li, J. A Comprehensive Survey on Design and Application of Autoencoder in Deep Learning. *Appl. Soft Comput.* **2023**, *138*, 110176, doi:10.1016/j.asoc.2023.110176.
96. Zhong, J.; Liu, Z.; Chen, X. Transformer-Based Models and Hardware Acceleration Analysis in Autonomous Driving: A Survey. 2023.
97. Ding, Z.; Sun, Y.; Xu, S.; Pan, Y.; Peng, Y.; Mao, Z. Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing. *Robotics* **2023**, *12*, 100, doi:10.3390/robotics12040100.
98. Liang, L.; Ma, H.; Zhao, L.; Xie, X.; Hua, C.; Zhang, M.; Zhang, Y. Vehicle Detection Algorithms for Autonomous Driving: A Review. *Sensors* **2024**, *24*, 3088, doi:10.3390/s24103088.
99. Zhu, M.; Gong, Y.; Tian, C.; Zhu, Z. A Systematic Survey of Transformer-Based 3D Object Detection for Autonomous Driving: Methods, Challenges and Trends. *Drones* **2024**, *8*, 412, doi:10.3390/drones8080412.
100. El Natour, G.; Bresson, G.; Trichet, R. Multi-Sensors System and Deep Learning Models for Object Tracking. *Sensors* **2023**, *23*, 7804, doi:10.3390/s23187804.
101. Vennerød, C.B.; Kjærran, A.; Bugge, E.S. Long Short-Term Memory RNN. 2021.
102. Taye, M.M. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers* **2023**, *12*, 91, doi:10.3390/computers12050091.
103. Reda, M.; Onsy, A.; Haikal, A.Y.; Ghanbari, A. Path Planning Algorithms in the Autonomous Driving System: A Comprehensive Review. *Robot. Auton. Syst.* **2024**, *174*, 104630, doi:10.1016/j.robot.2024.104630.
104. Jin, X.-B.; Chen, W.; Ma, H.-J.; Kong, J.-L.; Su, T.-L.; Bai, Y.-T. Parameter-Free State Estimation Based on Kalman Filter with Attention Learning for GPS Tracking in Autonomous Driving System. *Sensors* **2023**, *23*, 8650, doi:10.3390/s23208650.
105. Brownlee, J. A Gentle Introduction to Expectation-Maximization (EM Algorithm). *MachineLearningMastery.com* 2019.
106. Li, Y.; Li, S.; Du, H.; Chen, L.; Zhang, D.; Li, Y. YOLO-ACN: Focusing on Small Target and Occluded Object Detection. *IEEE Access* **2020**, *8*, 227288–227303, doi:10.1109/ACCESS.2020.3046515.
107. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. 2018.
108. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. 2016.
109. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS — Improving Object Detection with One Line of Code. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); October 2017; pp. 5562–5570.
110. Singh, A. Selecting the Right Bounding Box Using Non-Max Suppression (with Implementation). *Anal. Vidhya* 2020.
111. Wan, E.A.; Van Der Merwe, R. The Unscented Kalman Filter for Nonlinear Estimation. In Proceedings of the Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373); October 2000; pp. 153–158.
112. Hao, Y.; Xiong, Z.; Sun, F.; Wang, X. Comparison of Unscented Kalman Filters. In Proceedings of the 2007 International Conference on Mechatronics and Automation; August 2007; pp. 895–899.
113. Liu, D.; Zhang, J.; Jin, J.; Dai, Y.; Li, L. A New Approach of Obstacle Fusion Detection for Unmanned Surface Vehicle Using Dempster-Shafer Evidence Theory. *Appl. Ocean Res.* **2022**, *119*, 103016, doi:10.1016/j.apor.2021.103016.

114. Wibowo, A.; Trilaksono, B.R.; Hidayat, E.M.I.; Munir, R. Object Detection in Dense and Mixed Traffic for Autonomous Vehicles With Modified Yolo. *IEEE Access* **2023**, *11*, 134866–134877, doi:10.1109/ACCESS.2023.3335826.
115. Wong, C.-C.; Feng, H.-M.; Kuo, K.-L. Multi-Sensor Fusion Simultaneous Localization Mapping Based on Deep Reinforcement Learning and Multi-Model Adaptive Estimation. *Sensors* **2024**, *24*, 48, doi:10.3390/s24010048.
116. Charroud, A.; Moutaouakil, K.E.; Yahyaouy, A. Fast and Accurate Localization and Mapping Method for Self-Driving Vehicles Based on a Modified Clustering Particle Filter. *Multimed. Tools Appl.* **2023**, *82*, 18435–18457, doi:10.1007/s11042-022-14111-4.
117. Kang, D.; Kum, D. Camera and Radar Sensor Fusion for Robust Vehicle Localization via Vehicle Part Localization. *IEEE Access* **2020**, *8*, 75223–75236, doi:10.1109/ACCESS.2020.2985075.
118. Stroescu, A.; Daniel, L.; Gashinova, M. Combined Object Detection and Tracking on High Resolution Radar Imagery for Autonomous Driving Using Deep Neural Networks and Particle Filters. In Proceedings of the 2020 IEEE Radar Conference (RadarConf20); September 2020; pp. 1–6.
119. Kim, S.; Jang, M.; La, H.; Oh, K. Development of a Particle Filter-Based Path Tracking Algorithm of Autonomous Trucks with a Single Steering and Driving Module Using a Monocular Camera. *Sensors* **2023**, *23*, 3650, doi:10.3390/s23073650.
120. Kusenbach, M.; Luettel, T.; Wuensche, H.-J. Fast Object Classification for Autonomous Driving Using Shape and Motion Information Applying the Dempster-Shafer Theory. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC); September 2020; pp. 1–6.
121. Srivastava, R.P. Dempster-Shafer Theory of Belief Functions: A Language for Managing Uncertainties in the Real-World Problems. *Int. J. Finance Entrep. Sustain.* **2022**, *2*, doi:10.56763/ijfes.v1i.30.
122. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716, doi:10.3390/make5040083.
123. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149, doi:10.1109/TPAMI.2016.2577031.
124. Yang, W.; Li, Z.; Wang, C.; Li, J. A Multi-Task Faster R-CNN Method for 3D Vehicle Detection Based on a Single Image. *Appl. Soft Comput.* **2020**, *95*, 106533, doi:10.1016/j.asoc.2020.106533.
125. Cortés Gallardo Medina, E.; Velazquez Espitia, V.M.; Chípuli Silva, D.; Fernández Ruiz de las Cuevas, S.; Palacios Hirata, M.; Zhu Chen, A.; González González, J.Á.; Bustamante-Bello, R.; Moreno-García, C.F. Object Detection, Distributed Cloud Computing and Parallelization Techniques for Autonomous Driving Systems. *Appl. Sci.* **2021**, *11*, 2925, doi:10.3390/app11072925.
126. Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Chapman, M.A.; Cao, D.; Li, J. Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 3412–3432, doi:10.1109/TNNLS.2020.3015992.
127. Kang, D.; Wong, A.; Lee, B.; Kim, J. Real-Time Semantic Segmentation of 3D Point Cloud for Autonomous Driving. *Electronics* **2021**, *10*, 1960, doi:10.3390/electronics10161960.
128. Paigwar, A.; Erkent, Ö.; Sierra-Gonzalez, D.; Laugier, C. GndNet: Fast Ground Plane Estimation and Point Cloud Segmentation for Autonomous Vehicles. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); October 2020; pp. 2150–2156.
129. Lu, W.; Zhou, Y.; Wan, G.; Hou, S.; Song, S. L3-Net: Towards Learning Based LiDAR Localization for Autonomous Driving.; 2019; pp. 6389–6398.
130. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 722–739, doi:10.1109/TITS.2020.3023541.
131. Cheng, W.; Yin, J.; Li, W.; Yang, R.; Shen, J. Language-Guided 3D Object Detection in Point Cloud for Autonomous Driving 2023.
132. Alaba, S.Y.; Gurbuz, A.C.; Ball, J.E. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. *World Electr. Veh. J.* **2024**, *15*, 20, doi:10.3390/wevj15010020.

133. Wang, Y.; Mao, Q.; Zhu, H.; Deng, J.; Zhang, Y.; Ji, J.; Li, H.; Zhang, Y. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey 2023.
134. Mao, J.; Shi, S.; Wang, X.; Li, H. 3D Object Detection for Autonomous Driving: A Comprehensive Survey. *Int. J. Comput. Vis.* **2023**, *131*, 1909–1963, doi:10.1007/s11263-023-01790-1.
135. Zhu, M.; Gong, Y.; Tian, C.; Zhu, Z. A Systematic Survey of Transformer-Based 3D Object Detection for Autonomous Driving: Methods, Challenges and Trends. *Drones* **2024**, *8*, 412, doi:10.3390/drones8080412.
136. Wang, X.; Li, K.; Chehri, A. Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 1148–1165, doi:10.1109/TITS.2023.3317372.
137. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. 2024.
138. Qi, C.; Yi, L.; Su, H.; Guibas, L. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *ArXiv* **2017**.
139. Srivastav, A.; Mandal, S. Radars for Autonomous Driving: A Review of Deep Learning Methods and Challenges. *IEEE Access* **2023**, *11*, 97147–97168, doi:10.1109/ACCESS.2023.3312382.
140. Liu, C.; Liu, S.; Zhang, C.; Huang, Y.; Wang, H. Multipath Propagation Analysis and Ghost Target Removal for FMCW Automotive Radars. In Proceedings of the IET International Radar Conference (IET IRC 2020); November 2020; Vol. 2020, pp. 330–334.
141. Hasirlioglu, S.; Riener, A. Challenges in Object Detection Under Rainy Weather Conditions. In Proceedings of the Intelligent Transport Systems, From Research and Development to the Market Uptake; Ferreira, J.C., Martins, A.L., Monteiro, V., Eds.; Springer International Publishing: Cham, 2019; pp. 53–65.
142. Zhang, Y.; Liu, K.; Bao, H.; Qian, X.; Wang, Z.; Ye, S.; Wang, W. AFTR: A Robustness Multi-Sensor Fusion Model for 3D Object Detection Based on Adaptive Fusion Transformer. *Sensors* **2023**, *23*, 8400, doi:10.3390/s23208400.
143. Liu, W.; Zhu, J.; Zhuo, G.; Fu, W.; Meng, Z.; Lu, Y.; Hua, M.; Qiao, F.; Li, Y.; He, Y.; et al. UniMSF: A Unified Multi-Sensor Fusion Framework for Intelligent Transportation System Global Localization. 2024.
144. Dickert, C. Network Overload? Adding Up the Data Produced By Connected Cars. Available online: https://www.visualcapitalist.com/network-overload/ (accessed on 21 December 2024).
145. Vakulov, A. Addressing Data Processing Challenges in Autonomous Vehicles. Available online: https://www.iotforall.com/addressing-data-processing-challenges-in-autonomous-vehicles (accessed on 21 December 2024).
146. Griggs, T.; Wakabayashi, D. How a Self-Driving Uber Killed a Pedestrian in Arizona. *N. Y. Times* 2018.
147. Parekh, D.; Poddar, N.; Rajpurkar, A.; Chahal, M.; Kumar, N.; Joshi, G.P.; Cho, W. A Review on Autonomous Vehicles: Progress, Methods and Challenges. *Electronics* **2022**, *11*, 2162, doi:10.3390/electronics11142162.
148. Oh, C.; Yoon, J. Hardware Acceleration Technology for Deep-Learning in Autonomous Vehicles. In Proceedings of the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp); February 2019; pp. 1–3.
149. Singh, J. AI-Driven Path Planning in Autonomous Vehicles: Algorithms for Safe and Efficient Navigation in Dynamic Environments. *J. AI-Assist. Sci. Discov.* **2024**, *4*, 48–88.
150. Islayem, R.; Alhosani, F.; Hashem, R.; Alzaabi, A.; Meribout, M. Hardware Accelerators for Autonomous Cars: A Review. 2024.
151. Narain, S.; Ranganathan, A.; Noubir, G. Security of GPS/INS Based On-Road Location Tracking Systems. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP); May 2019; pp. 587–601.
152. Kloukiniotis, A.; Papandreou, A.; Lalos, A.; Kapsalas, P.; Nguyen, D.-V.; Moustakas, K. Countering Adversarial Attacks on Autonomous Vehicles Using Denoising Techniques: A Review. *IEEE Open J. Intell. Transp. Syst.* **2022**, *3*, 61–80, doi:10.1109/OJITS.2022.3142612.
153. Hataba, M.; Sherif, A.; Mahmoud, M.; Abdallah, M.; Alasmary, W. Security and Privacy Issues in Autonomous Vehicles: A Layer-Based Survey. *IEEE Open J. Commun. Soc.* **2022**, *3*, 811–829, doi:10.1109/OJCOMS.2022.3169500.
154. Hamad, M.; Steinhorst, S. Security Challenges in Autonomous Systems Design. 2023.
155. Yan, X.; Wang, H. Survey on Zero-Trust Network Security. In Proceedings of the Artificial Intelligence and Security; Sun, X., Wang, J., Bertino, E., Eds.; Springer: Singapore, 2020; pp. 50–60.

156. Pham, M.; Xiong, K. A Survey on Security Attacks and Defense Techniques for Connected and Autonomous Vehicles. *Comput. Secur.* **2021**, *109*, 102269, doi:10.1016/j.cose.2021.102269.
157. Girdhar, M.; Hong, J.; Moore, J. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models. *IEEE Open J. Veh. Technol.* **2023**, *4*, 417–437, doi:10.1109/OJVT.2023.3265363.
158. Giannaros, A.; Karras, A.; Theodorakopoulos, L.; Karras, C.; Kranias, P.; Schizas, N.; Kalogeratos, G.; Tsolis, D. Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions. *J. Cybersecurity Priv.* **2023**, *3*, 493–543, doi:10.3390/jcp3030025.
159. Bendiab, G.; Hameurlaine, A.; Germanos, G.; Kolokotronis, N.; Shiaeles, S. Autonomous Vehicles Security: Challenges and Solutions Using Blockchain and Artificial Intelligence. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 3614–3637, doi:10.1109/TITS.2023.3236274.
160. Wang, Z.; Wei, H.; Wang, J.; Zeng, X.; Chang, Y. Security Issues and Solutions for Connected and Autonomous Vehicles in a Sustainable City: A Survey. *Sustainability* **2022**, *14*, 12409, doi:10.3390/su141912409.
161. Ahmad, U.; Han, M.; Mahmood, S. Enhancing Security in Connected and Autonomous Vehicles: A Pairing Approach and Machine Learning Integration. *Appl. Sci.* **2024**, *14*, 5648, doi:10.3390/app14135648.
162. Zablocki, É.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges. *Int. J. Comput. Vis.* **2022**, *130*, 2425–2452, doi:10.1007/s11263-022-01657-x.
163. Nastjuk, I.; Herrenkind, B.; Marrone, M.; Brendel, A.B.; Kolbe, L.M. What Drives the Acceptance of Autonomous Driving? An Investigation of Acceptance Factors from an End-User's Perspective. *Technol. Forecast. Soc. Change* **2020**, *161*, 120319, doi:10.1016/j.techfore.2020.120319.
164. Lees, M.; Uys, W.; Oosterwyk, G.; Van Belle, J.-P. *Factors Influencing the User Acceptance of Autonomous Vehicle AI Technologies in South Africa*; 2022;
165. Mu, J.; Zhou, L.; Yang, C. Research on the Behavior Influence Mechanism of Users' Continuous Usage of Autonomous Driving Systems Based on the Extended Technology Acceptance Model and External Factors. *Sustainability* **2024**, *16*, 9696, doi:10.3390/su16229696.
166. Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artif. Intell.* **2021**, *296*, 103473, doi:10.1016/j.artint.2021.103473.
167. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput Surv* **2023**, *55*, 194:1-194:33, doi:10.1145/3561048.
168. Hussain, F.; Hussain, R.; Hossain, E. Explainable Artificial Intelligence (XAI): An Engineering Perspective 2021.
169. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115, doi:10.1016/j.inffus.2019.12.012.
170. Kuznietsov, A.; Gyevnar, B.; Wang, C.; Peters, S.; Albrecht, S.V. Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 19342–19364, doi:10.1109/TITS.2024.3474469.
171. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805, doi:10.1016/j.inffus.2023.101805.
172. Napkin AI - The Visual AI for Business Storytelling. Available online: https://wwww.napkin.ai (accessed on 27 December 2024).
173. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215, doi:10.1038/s42256-019-0048-x.

174. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput Surv* **2018**, *51*, 93:1-93:42, doi:10.1145/3236009.
175. Derrick, A. The 4 Key Principles of Explainable AI Applications. Available online: https://virtualitics.com/the-4-key-principles-of-explainable-ai-applications/ (accessed on 26 December 2024).
176. *Optimus - Gen 2 | Tesla*; 2023;
177. Biba, J.; Urwin, M. Tesla's Robot, Optimus: Everything We Know Available online: https://builtin.com/robotics/tesla-robot (accessed on 27 December 2024).
178. LeNail, A. NN-SVG: Publication-Ready Neural Network Architecture Schematics. *J. Open Source Softw.* **2019**, *4*, 747, doi:10.21105/joss.00747.
179. Xiong, H.; Li, X.; Zhang, X.; Chen, J.; Sun, X.; Li, Y.; Sun, Z.; Du, M. Towards Explainable Artificial Intelligence (XAI): A Data Mining Perspective 2024.
180. Wang, S.; Zhou, T.; Bilmes, J. Bias Also Matters: Bias Attribution for Deep Neural Network Explanation.; May 24 2019.
181. Perteneder, F. Understanding Black-Box ML Models with Explainable AI Available online: https://www.dynatrace.com/news/blog/explainable-ai/ (accessed on 29 December 2024).
182. Buhrmester, V.; Münch, D.; Arens, M. Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 966–989, doi:10.3390/make3040048.
183. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.* **2024**, *16*, 45–74, doi:10.1007/s12559-023-10179-8.
184. Xu, B.; Yang, G. Interpretability Research of Deep Learning: A Literature Survey. *Inf. Fusion* **2025**, *115*, 102721, doi:10.1016/j.inffus.2024.102721.
185. Lyssenko, M.; Pimplikar, P.; Bieshaar, M.; Nozarian, F.; Triebel, R. A Safety-Adapted Loss for Pedestrian Detection in Automated Driving 2024.
186. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113, doi:10.1109/ACCESS.2019.2949286.
187. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Interpretable Machine Learning: Definitions, Methods, and Applications. *Proc. Natl. Acad. Sci.* **2019**, *116*, 22071–22080, doi:10.1073/pnas.1900654116.
188. Molnar, C. *8.4 Functional Decomposition | Interpretable Machine Learning*; 2024; ISBN 978-0-244-76852-2.
189. Kawaguchi, K. Deep Learning without Poor Local Minima 2016.
190. Datta, A.; Sen, S.; Zick, Y. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP); May 2016; pp. 598–617.
191. Liu, D.Y. Explainable AI Techniques for Transparency in Autonomous Vehicle Decision-Making. *J. AI Healthc. Med.* **2023**, *3*, 114–134.
192. Chaudhary, G. Unveiling the Black Box: Bringing Algorithmic Transparency to AI. *Masaryk Univ. J. Law Technol.* **2024**, *18*, 93–122.
193. Alicioglu, G.; Sun, B. A Survey of Visual Analytics for Explainable Artificial Intelligence Methods. *Comput. Graph.* **2022**, *102*, 502–520, doi:10.1016/j.cag.2021.09.002.
194. Rjoub, G.; Bentahar, J.; Wahab, O.A.; Mizouni, R.; Song, A.; Cohen, R.; Otrok, H.; Mourad, A. A Survey on Explainable Artificial Intelligence for Cybersecurity. *IEEE Trans. Netw. Serv. Manag.* **2023**, *20*, 5115–5140, doi:10.1109/TNSM.2023.3282740.
195. Basheer, K.C.S. Understanding Generalized Additive Models (GAMs): A Comprehensive Guide. *Anal. Vidhya* 2023.
196. Espinoza, J.; Delpiano, R. Statistical Models of Interactions between Vehicles during Overtaking Maneuvers. *Transp. Res. Rec. J. Transp. Res. Board* **2023**, *2678*, doi:10.1177/03611981231184230.
197. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization.; 2017; pp. 618–626.

198. Yamauchi, T.; Ishikawa, M. Spatial Sensitive GRAD-CAM: Visual Explanations for Object Detection by Incorporating Spatial Sensitivity. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP); October 2022; pp. 256–260.
199. Quattrocchio, L. Integration of Uncertainty into Explainability Methods to Enhance AI Transparency in Brain MRI Classification. laurea, Politecnico di Torino, 2024.
200. Letham, B.; Rudin, C.; McCormick, T.H.; Madigan, D. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *Ann. Appl. Stat.* **2015**, *9*, 1350–1371, doi:10.1214/15-AOAS848.
201. Rjoub, G.; Bentahar, J.; Wahab, O.A. Explainable Trust-Aware Selection of Autonomous Vehicles Using LIME for One-Shot Federated Learning. In Proceedings of the 2023 International Wireless Communications and Mobile Computing (IWCMC); June 2023; pp. 524–529.
202. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160, doi:10.1109/ACCESS.2018.2870052.
203. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems; Curran Associates Inc.: Red Hook, NY, USA, December 4 2017; pp. 4768–4777.
204. Onyekpe, U.; Lu, Y.; Apostolopoulou, E.; Palade, V.; Eyo, E.U.; Kanarachos, S. Explainable Machine Learning for Autonomous Vehicle Positioning Using SHAP. In *Explainable AI: Foundations, Methodologies and Applications*; Mehta, M., Palade, V., Chatterjee, I., Eds.; Springer International Publishing: Cham, 2023; pp. 157–183 ISBN 978-3-031-12807-3.
205. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. 2014.
206. Karatsiolis, S.; Kamilaris, A. A Model-Agnostic Approach for Generating Saliency Maps to Explain Inferred Decisions of Deep Learning Models. 2022.
207. Cooper, J.; Arandjelović, O.; Harrison, D.J. Believe the HiPe: Hierarchical Perturbation for Fast, Robust, and Model-Agnostic Saliency Mapping. *Pattern Recognit.* **2022**, *129*, 108743, doi:10.1016/j.patcog.2022.108743.
208. Yang, S.; Berdine, G. Interpretable Artificial Intelligence (AI) – Saliency Maps. *Southwest Respir. Crit. Care Chron.* **2023**, *11*, 31–37, doi:10.12746/swrccc.v11i48.1209.
209. Ding, N.; Zhang, C.; Eskandarian, A. SalienDet: A Saliency-Based Feature Enhancement Algorithm for Object Detection for Autonomous Driving. *IEEE Trans. Intell. Veh.* **2024**, *9*, 2624–2635, doi:10.1109/TIV.2023.3287359.
210. Salih, A.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Menegaz, G.; Lekadir, K. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2024**, 2400304, doi:10.1002/aisy.202400304.
211. Dieber, J.; Kirrane, S. Why Model Why? Assessing the Strengths and Limitations of LIME. 2020.
212. Tursunalieva, A.; Alexander, D.L.J.; Dunne, R.; Li, J.; Riera, L.; Zhao, Y. Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. *Appl. Sci.* **2024**, *14*, 496, doi:10.3390/app14020496.
213. Retzlaff, C.O.; Angerschmid, A.; Saranti, A.; Schneeberger, D.; Röttger, R.; Müller, H.; Holzinger, A. Post-Hoc vs Ante-Hoc Explanations: xAI Design Guidelines for Data Scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243, doi:10.1016/j.cogsys.2024.101243.
214. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18, doi:10.3390/e23010018.
215. Tahir, H.A.; Alayed, W.; Hassan, W.U.; Haider, A. A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration. *Sensors* **2024**, *24*, 6776, doi:10.3390/s24216776.
216. Ortigossa, E.S.; Gonçalves, T.; Nonato, L.G. EXplainable Artificial Intelligence (XAI)—From Theory to Methods and Applications. *IEEE Access* **2024**, *12*, 80799–80846, doi:10.1109/ACCESS.2024.3409843.
217. Basheer, K.C.S. Understanding Generalized Additive Models (GAMs): A Comprehensive Guide. *Anal. Vidhya* 2023.

218. Y, S.; Challa, M. A Comparative Analysis of Explainable AI Techniques for Enhanced Model Interpretability. In Proceedings of the 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN); June 2023; pp. 229–234.
219. Vignesh Everything You Need to Know about LIME. *Anal. Vidhya* 2022.
220. Santos, M.R.; Guedes, A.; Sanchez-Gendriz, I. SHapley Additive exPlanations (SHAP) for Efficient Feature Selection in Rolling Bearing Fault Diagnosis. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 316–341, doi:10.3390/make6010016.
221. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; Association for Computing Machinery: New York, NY, USA, February 7 2020; pp. 180–186.
222. Ghorbani, A.; Abid, A.; Zou, J. Interpretation of Neural Networks Is Fragile. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 3681–3688, doi:10.1609/aaai.v33i01.33013681.
223. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); March 2018; pp. 839–847.
224. Ahern, I.; Noack, A.; Guzman-Nateras, L.; Dou, D.; Li, B.; Huan, J. NormLime: A New Feature Importance Metric for Explaining Deep Neural Networks. 2019.
225. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, doi:10.1609/aaai.v32i1.11491.
226. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognit.* **2017**, *65*, 211–222, doi:10.1016/j.patcog.2016.11.008.
227. Huang, J.; Wang, Z.; Li, D.; Liu, Y. The Analysis and Development of an XAI Process on Feature Contribution Explanation. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data); December 2022; pp. 5039–5048.
228. Agarwal, G. Explainable AI (XAI): Permutation Feature Importance. *Medium* 2022.
229. Majella XAI Problems Part 1 of 3: Feature Importances Are Not Good Enough. *Elula* 2021.
230. Cambria, E.; Malandri, L.; Mercorio, F.; Mezzanzanica, M.; Nobani, N. A Survey on XAI and Natural Language Explanations. *Inf. Process. Manag.* **2023**, *60*, 103111, doi:10.1016/j.ipm.2022.103111.
231. Andres, A.; Martinez-Seras, A.; Laña, I.; Del Ser, J. On the Black-Box Explainability of Object Detection Models for Safe and Trustworthy Industrial Applications. *Results Eng.* **2024**, *24*, 103498, doi:10.1016/j.rineng.2024.103498.
232. Shylenok, D.Y. Explainable AI for Transparent Decision-Making in Autonomous Vehicle Systems. *Afr. J. Artif. Intell. Sustain. Dev.* **2023**, *3*, 320–341.
233. Moradi, M.; Yan, K.; Colwell, D.; Samwald, M.; Asgari, R. Model-Agnostic Explainable Artificial Intelligence for Object Detection in Image Data. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109183, doi:10.1016/j.engappai.2024.109183.
234. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015, doi:10.1109/ACCESS.2023.3294569.
235. Greifenstein, M. Factors Influencing the User Behaviour of Shared Autonomous Vehicles (SAVs): A Systematic Literature Review. *Transp. Res. Part F Traffic Psychol. Behav.* **2024**, *100*, 323–345, doi:10.1016/j.trf.2023.10.027.
236. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.* **2020**, 1–15, doi:10.1145/3313831.3376590.
237. Shin, D. The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102551, doi:10.1016/j.ijhcs.2020.102551.
238. SmythOS - Explainable AI in Autonomous Vehicles: Building Transparency and Trust on the Road. 2024.
239. Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 11 January 2025).
240. Nasim, M.A.A.; Biswas, P.; Rashid, A.; Biswas, A.; Gupta, K.D. Trustworthy XAI and Application 2024.

241. Gunning, D.; Vorm, E.; Wang, Y.; Turek, M. DARPA's Explainable AI (XAI) Program: A Retrospective. 2021.
242. Cugny, R.; Aligon, J.; Chevalier, M.; Roman Jimenez, G.; Teste, O. AutoXAI: A Framework to Automatically Select the Most Adapted XAI Solution. In Proceedings of the Proceedings of the 31st ACM International Conference on Information & Knowledge Management; Association for Computing Machinery: New York, NY, USA, October 17 2022; pp. 315–324.
243. Gadekallu, T.R.; Maddikunta, P.K.R.; Boopathy, P.; Deepa, N.; Chengoden, R.; Victor, N.; Wang, W.; Wang, W.; Zhu, Y.; Dev, K. XAI for Industry 5.0 -Concepts, Opportunities, Challenges and Future Directions. *IEEE Open J. Commun. Soc.* **2024**, 1–1, doi:10.1109/OJCOMS.2024.3473891.
244. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.-W.; Newman, S.-F.; Kim, J.; et al. Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760, doi:10.1038/s41551-018-0304-0.
245. Holzinger, A.; Malle, B.; Saranti, A.; Pfeifer, B. Towards Multi-Modal Causability with Graph Neural Networks Enabling Information Fusion for Explainable AI. *Inf. Fusion* **2021**, *71*, 28–37, doi:10.1016/j.inffus.2021.01.008.
246. European Approach to Artificial Intelligence | Shaping Europe's Digital Future Available online: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed on 14 January 2025).
247. Chen, L.; Sinavski, O.; Hünermann, J.; Karnsund, A.; Willmott, A.J.; Birch, D.; Maund, D.; Shotton, J. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving 2023.
248. Greyling, C. Using LLMs For Autonomous Vehicles. *Medium* 2024.
249. Mankodiya, H.; Obaidat, M.S.; Gupta, R.; Tanwar, S. XAI-AV: Explainable Artificial Intelligence for Trust Management in Autonomous Vehicles. In Proceedings of the 2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI); October 2021; pp. 1–5.
250. Salehi, J. Explainable AI for Real-Time Threat Analysis in Autonomous Vehicle Networks. *Afr. J. Artif. Intell. Sustain. Dev.* **2023**, *3*, 294–315.