applied sciences

MDPI

Review

Image Analysis in Autonomous Vehicles: A Review of the Latest AI Solutions and Their Comparison

Michał Kozłowski , Szymon Racewicz and Sławomir Wierzbicki *

Faculty of Technical Sciences, University of Warmia and Mazury in Olsztyn, 11 Oczapowskiego Str., 10-719 Olsztyn, Poland; michal.kozlowski@uwm.edu.pl (M.K.); szymon.racewicz@uwm.edu.pl (S.R.)
* Correspondence: slawekw@uwm.edu.pl

Abstract: The integration of advanced image analysis using artificial intelligence (AI) is pivotal for the evolution of autonomous vehicles (AVs). This article provides a thorough review of the most significant datasets and latest state-of-the-art AI solutions employed in image analysis for AVs. Datasets such as Cityscapes, NuScenes, CARLA, and Talk2Car form the benchmarks for training and evaluating different AI models, with unique characteristics catering to various aspects of autonomous driving. Key AI methodologies, including Convolutional Neural Networks (CNNs), Transformer models, Generative Adversarial Networks (GANs), and Vision Language Models (VLMs), are discussed. The article also presents a comparative analysis of various AI techniques in real-world scenarios, focusing on semantic image segmentation, 3D object detection, vehicle control in virtual environments, and vehicle interaction using natural language. Simultaneously, the roles of multisensor datasets and simulation platforms like AirSim, TORCS, and SUMMIT in enriching the training data and testing environments for AVs are highlighted. By synthesizing information on datasets, AI solutions, and comparative performance evaluations, this article serves as a crucial resource for researchers, developers, and industry stakeholders, offering a clear view of the current landscape and future directions in autonomous vehicle image analysis technologies.

Keywords: autonomous vehicles; image analysis; AI solutions; safety features

Citation: Kozłowski, M.; Racewicz, S.; Wierzbicki, S. Image Analysis in Autonomous Vehicles: A Review of the Latest AI Solutions and Their Comparison. *Appl. Sci.* **2024**, *14*, 8150. https://doi.org/10.3390/app14188150

Academic Editors: Ali Zemouche, Karim Hammoudi, Rémi Boutteau and Redouane Khemmar

Received: 20 July 2024
Revised: 30 August 2024
Accepted: 9 September 2024
Published: 11 September 2024

1. Introduction

Environmental image analysis is pivotal in autonomous vehicle development and operation, serving as the cornerstone of their perception and decision-making capabilities. Accurate and real-time image analysis allows these vehicles to interpret their surroundings effectively, facilitating safe and efficient navigation. A wide range of autonomous vehicles, including cars [1–3], ships [4,5], underwater vehicles [6–8], unmanned aerial vehicles (UAVs) [9–11], and robots [12,13], employ an array of sensors, including cameras, LiDAR, and radar, to capture detailed images and environmental information [4,10,11,14,15]. These images undergo sophisticated processing through algorithms and machine learning techniques [16,17] to detect, classify, and track objects such as pedestrians, cyclists, other vehicles, traffic signs, and lane markings.

Recently, Vision Language Models (VLMs) [18–20] have shown promise by integrating visual and textual information, enhancing the contextual understanding of environments, which is critical for real-time processing and decision making in autonomous systems. Additionally, federated learning allows for decentralized model training, enhancing privacy while leveraging data from multiple sources, thus refining image analysis algorithms. Federated learning is also possible through the use of blockchain technology, i.e., Internet Computer [21]. Due to the decentralized nature of blockchain technology, data can be stored and processed in a distributed manner, eliminating the need for a central server that could become a target for attacks or privacy violations [22]. Internet Computer enables the creation of smart contracts that can manage the model training process in a secure and transparent manner [23]. Additionally, this technology can be integrated with federated learning, where models are trained locally on user devices and only updated weights are sent to the blockchain network. This allows users' data to remain on their devices, minimizing the risk of data leakage.

Autonomous vehicles are categorized into different levels of automation, as defined by the Society of Automotive Engineers (SAE) (Figure 1), ranging from level 0 (no automation) to level 5 (full automation) [24]. Higher levels indicate increased system independence and decreased human intervention, primarily driven by advancements in image analysis technology. At level 0, image analysis is not used and the driver is responsible for all aspects of driving. At levels 1 (Driver Assistance) and 2 (Partial Automation), image analysis becomes more critical, managing tasks such as adaptive cruise control, lane-keeping assistance, and semi-automated features like auto-steering and traffic-aware cruise control, respectively. Level 3 (Conditional Automation) marks a significant shift, where vehicles can take full control under certain conditions, heavily relying on high-definition cameras and advanced algorithms for monitoring the environment and detecting objects. At level 4 (High Automation), vehicles operate without human intervention in most scenarios, necessitating robust image analysis systems that integrate visual data with inputs from other sensors such as LiDAR and radar. Finally, level 5 (Full Automation) epitomizes vehicle autonomy, with the vehicle handling all driving aspects in any condition without human oversight, requiring image analysis capabilities surpassing human abilities in interpreting visual information and making instantaneous decisions

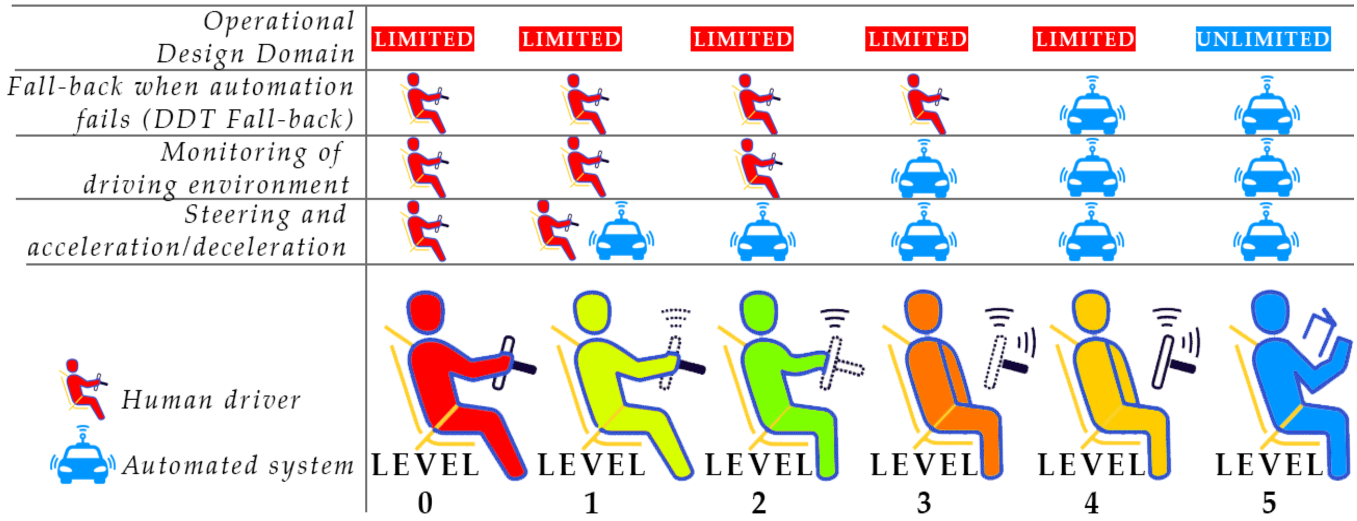
Operational Design Domain
LIMITED
LIMITED
LIMITED
LIMITED
LIMITED
UNLIMITED
Fall-back when automation fails (DDT Fall-back)
Monitoring of driving environment
Steering and acceleration/deceleration
Human driver
Automated system
LEVEL 0
LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
LEVEL 5

Figure 1. SAE J3016 Levels of Driving Automation.

The significance of image analysis spans all autonomy levels, enhancing situational awareness, safety and redundancy, complex decision making, adaptability, and precise mapping and localization, ultimately supporting the regulatory and social acceptance of autonomous systems. Technologies like federated learning facilitate collective model training from distributed vehicle fleets, improving the robustness and generalizability of image analysis algorithms without compromising data privacy.

Image analysis techniques in autonomous vehicles are employed across various facets. For example, computer vision analysis is crucial for vehicular safety applications, including collision avoidance and 3D map building [25,26]. Real-time image processing methods are designed for road detection, obstacle recognition, traffic light detection, and even number plate recognition [27–29]. Road detection techniques utilize similarity analysis and structural similarity index analysis to identify road surfaces in grid image areas [30]. Object detection and tracking remain challenging tasks, with image classification algorithms such as CNNs playing a pivotal role in steering vehicles [31–33].

Environmental image analysis is essential for numerous reasons. Primarily, it ensures safety by enabling vehicles to recognize and respond to dynamic and static objects in their path, thus preventing collisions. It enhances situational awareness, allowing vehicles to make informed decisions regarding speed, direction, and maneuvering based on the current context, such as traffic conditions and road layout.

Despite significant progress, several challenges remain. These include the need for more sophisticated object detection and tracking algorithms, especially under adverse conditions such as poor lighting, inclement weather [34,35], and complex urban landscapes. Moreover, the effectiveness of image analysis techniques in ensuring situational awareness, adaptability, and efficiency must continue to evolve in tandem with technological advancements.

Moreover, the proliferation of diverse datasets and evolving benchmarks necessitates a comprehensive review of existing datasets and a comparison of the performances of leading models. This review uniquely aims to fill this gap by systematically comparing available datasets and evaluating key models in terms of their efficacy in diverse real-world scenarios. This includes exploring how these models fare in complex tasks such as 3D map building, real-time road and obstacle detection, and the behavior prediction of other road users.

Environmental image analysis ensures the safety, situational awareness, adaptability, and operational efficiency of autonomous vehicles. This review aims to systematically compare and evaluate the performances of current datasets and models, providing a novel perspective that is essential for guiding future advancements in this field.

2. AI in Autonomous Vehicles

Research on image analysis for autonomous vehicles is an expansive and swiftly evolving domain intersecting with various interdisciplinary areas such as vision, machine learning, sensor fusion, federated learning, and more. Key advancements in recent years include the application of cutting-edge methodologies such as Convolutional Neural Networks (CNNs), transformer models, Generative Adversarial Networks (GANs), and Vision Language Models (VLMs). The existing literature and publications cover a wide range of topics, ranging from algorithm development to practical applications and system validation. Object detection and classification are fundamental tasks for autonomous vehicles, enabling the identification of pedestrians, vehicles, traffic signs, and other road users [36]. Notable research includes YOLO (You Only Look Once), proposed by Redmon et al., known for its efficiency and accuracy in real-time object detection [37,38]. In the latest YOLO model versions, there have been significant improvements in detecting small objects in images. Advanced image processing techniques and optimized neural network architectures have enabled these models to achieve a greater precision and effectiveness in identifying small elements, marking a substantial advancement in the field of computer vision [39–41]. Additionally, R-CNN (Region-Based Convolutional Neural Network) [42], developed by Ross Girshick and colleagues, integrates region proposal networks with deep learning, significantly improving detection performance. Semantic segmentation involves classifying each pixel in an image to understand the full scene context. Exemplary works include DeepLab developed at Google Research, which uses deep convolutional networks and atrous convolutions for detailed spatial hierarchies [43], and SegNet [44] proposed by Badrinarayanan et al., designed for pixel-wise semantic segmentation while conserving memory and computational resources, making it practical for real-time applications. The continuous development of better image analysis algorithms and architectures is paramount. Notable advancements include EfficientNet by the Google Brain team [45], which uses a compound scaling method for state-of-the-art results regarding efficiency and accuracy, and Vision Transformers (ViTs), which demonstrate promise in scaling and capturing long-range dependencies [46].

Depth estimation and stereo vision are crucial for understanding spatial relationships. Notable advances include traditional stereo matching techniques like Semi-Global

Matching (SGM) [47] by Hirschmüller, used in stereo vision systems for calculating depth maps, and monocular depth estimation using deep learning by Eigen et al. [48], which significantly advances the capability for depth estimation from single images. Sensor fusion combines data from multiple sensor types such as cameras, LiDAR, radar, and Vision Language Models (VLMs) to enhance perception system robustness and contextual understanding [18].

VLMs, which integrate visual data with textual information, are emerging as crucial tools for enhancing the contextual understanding of scenes in autonomous driving. By combining data streams from diverse sensor types—including cameras and LiDAR—VLMs facilitate improved perception system robustness, as they can leverage descriptive labels to enhance image analysis. Recent advancements, such as the DriveGenVLM framework developed by researchers at Columbia University, demonstrate the potential of VLMs in generating and analyzing real-world video sequences using diffusion models [49]. Additionally, a comprehensive review from 2024 highlights the applications of VLMs in various aspects of autonomous driving, including perception, navigation, planning, decision making, and control, emphasizing their role in improving safety and efficiency [50]. Furthermore, studies on lightweight and efficient multi-modal models (MMLMs) have revealed their capability to provide interpretable textual justifications and responses to safety-related tasks by utilizing images of road scenes and other data sources [51]. These developments underscore the significant potential of VLMs in enhancing the understanding and analysis of complex road environments, thereby advancing the field of autonomous driving.

Advances in deep learning for sensor fusion have been evaluated systematically in projects like KITTI [26,52], which combined visual data from cameras with depth information from LiDAR, and end-to-end learning approaches proposed by Chen et al. Ensuring the robustness of image analysis systems against adversarial attacks and overall safety is a critical research area. For instance, Goodfellow et al.'s seminal paper "Explaining and Harnessing Adversarial Examples" [53] discusses neural networks' vulnerability to small, intentionally crafted perturbations and proposes strategies for improving robustness. Safety and verification platforms like the CARLA [54] and TORCS [55] simulators provide rigorous testing and validation environments.

Generative Adversarial Networks (GANs), introduced by Ian Goodfellow in 2014, have revolutionized image analysis, particularly in autonomous vehicles [56]. GANs consist of a generator that creates realistic data and a discriminator that distinguishes between real and generated data. This process enables GANs to produce highly realistic images, enhancing training datasets for tasks like object detection and semantic segmentation, thus improving model robustness and generalizability [57–59]. GANs simulate diverse driving conditions, generating images of various weather scenarios, lighting conditions, and road environments, which broadens the training spectrum without the need for extensive real-world data collection. Models like SRGAN (Super-Resolution GAN) enhance image resolution, which is crucial for accurate object detection and scene understanding. Additionally, GANs reduce the noise in low-light images, improving their quality [60]. Techniques such as CycleGAN translate images between domains (e.g., synthetic to real world), essential for training models in simulated environments. Style transfer via GANs allows for models trained on one type of imagery (e.g., daytime) to perform well in another (e.g., nighttime) [61]. GANs also improve semantic segmentation accuracy. Models like SegAN use adversarial training for fine-grained results. Conditional GANs (cGANs) generate images based on specific inputs, aiding in training and validating segmentation models [62]. GANs create realistic virtual driving scenarios for testing autonomous driving algorithms, safely simulating rare and dangerous situations. They generate adversarial examples to test and enhance perception systems' robustness and predict future frames in video sequences, aiding in object motion understanding and navigation [63].

The incorporation of federated learning into autonomous vehicles represents a groundbreaking approach to maintaining data privacy and security. Federated learning enables models to be trained across decentralized data sources without centrally aggregating the data. This method ensures that sensitive data, essential for training AI models, remain on local devices without compromising privacy. Moreover, it allows for continual learning from diverse driving experiences, thereby enriching the models' generalizability and robustness [22].

Addressing the unique challenges of perception in adverse weather conditions is critical for reliable autonomous vehicle operation. Techniques must account for varying visibility, sensor noise, and environment-induced variability. Advanced methods in accident prediction leverage historical data, real-time sensor inputs, and predictive modeling. By analyzing patterns and signals, these models can foresee potential collision scenarios and initiate pre-emptive actions. Beyond ideal operating conditions, learning in non-standard situations such as construction zones, unexpected road hazards, and temporary traffic patterns requires adaptive algorithms capable of dynamic responses and reconfiguration.

In summary, the corpus of research and publications on image analysis in autonomous vehicles is expansive, including various critical areas spanning from fundamental algorithm development to practical deployments, ensuring safety, reliability, and widespread acceptance. Addressing the ongoing challenges in image analysis through continuous research and development is essential for advancing autonomous vehicle technology. The integration of sophisticated AI techniques, particularly deep learning, GANs, VLMs, and federated learning, continues to push the boundaries of what is possible in vehicle perception systems, paving the way for safer, more efficient, and intelligent autonomous transportation.

3. Overview of Available Datasets

The development and deployment of autonomous vehicles heavily rely on vast and comprehensive datasets to train, validate, and test their perception, decision-making, and control systems. These datasets encompass various forms of sensor data, including images, LiDAR point clouds, radar echoes, and GPS coordinates, capturing diverse driving environments, conditions, and scenarios, enabling the development of robust AI models for safe and effective autonomous driving. Image datasets are essential for tasks such as object detection, classification, semantic segmentation, and lane detection. The most used datasets include Cityscapes for understanding urban scenes, nuScenes for providing data for spatial object detection, and CARLA for testing models in realistic simulations. Radar datasets, though less common, provide valuable information on object speed and distance, exemplified by nuScenes, which includes radar data for comprehensive perception tasks. Multisensor datasets like Argoverse [64] offer synchronized data from cameras, LiDAR, radar, and GPS, enhancing sensor fusion techniques. Simulated datasets from platforms like CARLA and TORCS offer cost-effective and controlled data collection in diverse virtual environments, essential for training, planning, and control tasks. These datasets are crucial for training AI models by providing diverse examples, enabling rigorous validation and testing, offering standardized benchmarks, handling edge cases, and advancing research through rapid iteration. The proliferation of such datasets has accelerated research, demonstrated by thousands of papers (Tables 1–3) on applications like object detection, semantic segmentation, sensor fusion, and adversarial robustness, making datasets the cornerstone of autonomous vehicle development and innovation.

Table 1. Available image datasets that can be used to train deep neural models for implementation in autonomous vehicles.

Name	Description	Sensors Used	Samples Number	Classes Number	2D/3D	Cited	Ref.
Cityscapes	Provides semantic, instance-wise, and dense pixel annotations for 30 classes grouped into 8 categories.	Camera	25,000	30	2D&3D	3411	[65]
Waymo	High-resolution sensor data collected by Waymo Driver in various conditions.	Camera, LiDAR	105,000+	4	3D	398	[66]

Table 1. *Cont.*

Name	Description	Sensors Used	Samples Number	Classes Number	2D/3D	Cited	Ref.
SODA-D	High-quality images under driving scenarios for small object detection.	Camera	24,828	9	2D	232	[67]
KITTI-360	Popular KITTI dataset with comprehensive semantic/instance labels in 2D and 3D.	Camera, LiDAR	320,000+	19	2D&3D	181	[52]
IDD	Road scene understanding in unstructured environments dataset.	Camera	10,000+	34	2D&3D	90	[68]
INTERACTION	Contains naturalistic motions of traffic participants in highly interactive scenarios.	Traffic camera	35,000+	4	3D	73	[69]
SemanticPOSS	Three-dimensional semantic segmentation dataset collected in Peking University.	LiDAR	2988	14	3D	60	[70]
WoodScape	Extensive fisheye camera automotive dataset with nine tasks and 40 class annotations.	Fisheye camera	100,000+	40	2D&3D	49	[71]
Lost and Found	Lost-cargo image sequence dataset with pixel-wise annotations of obstacles and free space.	Stereo camera	2104	14	2D	47	[72]
DrivingStereo	Over 180k images for stereo vision, larger than KITTI Stereo dataset.	Stereo camera	182,000	10	3D	42	[73]
Fishyscapes	Evaluates pixel-wise uncertainty estimates towards detecting anomalous objects.	Camera	30,000	n/a	2D	44	[74]
ROAD Anomaly	Contains images of unusual dangers encountered by vehicles, such as animals and traffic cones.	Camera	1000	n/a	2D	44	[75]
PandaSet	Dataset captured with high-precision autonomous vehicle sensor kit.	Camera LiDAR	8000	32	3D	39	[76]
KITTI Road	Road and lane estimation benchmark.	Stereo camera, LiDAR	289	3	2D	38	[77]
MVSEC	Data collection designed for developing 3D perception algorithms for event-based cameras.	Event-based camera and stereo camera	4	n/a	3D	26	[78]
KAIST Urban	Raw sensor data for vehicle navigation with development tools in the ROS environment.	Camera and LiDAR	95,000	25	2D&3D	19	[79]
Cityscapes 3D	Extends the original Cityscapes dataset with 3D bounding box annotations for vehicles.	Stereo camera	5000	30	3D	10	[65]
RailSem19	Dataset for semantic rail scene understanding with images from rail vehicles.	Camera	850	19	2D	8	[80]
RoadAnomaly21	Contains images with at least one anomalous object such as animals or unknown vehicles.	Camera	1000	n/a	2D	8	[81]
EuroCity Persons	Annotations of pedestrians, cyclists, and riders in urban traffic scenes from 31 cities in Europe.	Camera	47,300	2	2D	6	[82]
DOLPHINS	Dataset for testing vehicle-to-everything (V2X) network in autonomous driving.	Camera, LiDAR	42,376	2	2D&3D	5	[83]
PSI	Dataset capturing dynamic intent changes for pedestrians crossing in front of ego vehicles.	Camera	1000	3	2D&3D	5	[84,85]
TICaM	Dataset for vehicle interior monitoring using a wide-angle depth camera.	Camera	1000	5	2D	5	[86]
Zenseact	Dataset collected over 2 years across 14 European countries with full sensor suite.	Camera, GNSS/IMU, LiDAR	100,000	5	2D&3D	45	[87]
OoDIS	Dataset for anomaly instance segmentation in autonomous driving.	Camera	4218	8	2D	4	[88]
LOOK	Real-world scenarios for autonomous vehicles focusing on pedestrian interactions.	Camera	16,000	18	2D	3	[89]

Table 2. Available video datasets that can be used to train deep neural models for implementation in autonomous vehicles.

Name	Description	Sensors Used	Samples Number	Classes Number	2D/3D	Cited	Ref.
nuScenes	Full autonomous vehicle data suite: 32-beam LiDAR, 6 cameras, and radars with complete 360° coverage.	6 cameras, 1 LIDAR, 5 RADAR, GPS, IMU	1.4 M	32	2D and 3D	1695	[90]
Virtual KITTI	Photo-realistic synthetic video dataset for several video understanding tasks.	4 cameras, LIDAR, RADAR, GPS, IMU	21,200	7	2D and 3D	124	[91]
Talk2Car	Cross-disciplinary dataset for grounding natural language into visual space.	Cameras, LiDAR	80,000	5	2D and 3D	105	[92]
CULane	Lane detection dataset collected by cameras mounted on six different vehicles in Beijing.	Cameras	133,000	6	2D	77	[93]
ApolloScape	Large dataset with over 140,000 video frames from various locations in China.	Cameras, LiDAR	140,000	18	2D and 3D	68	[94]
ROAD	Tests an autonomous vehicle's ability to detect road events using annotated videos.	Cameras, LiDAR	8000	4	2D	21	[95]
V2V4Real	Data collected by two vehicles equipped with multi-modal sensors driving together through diverse scenarios.	Cameras, LiDAR	20,000	3	2D	17	[96]
TITAN	700 labeled video-clips with odometry captured from a moving vehicle in Tokyo.	Cameras, LiDAR	10,000	5	2D and 3D	12	[97]
BrnoCompSpeed	Vehicles annotated with precise speed measurements from LiDAR and GPS tracks.	Cameras	38,000	4	2D	11	[98]
CCD	Real traffic accident videos captured by dashcams with diverse annotations.	Cameras, LiDAR, GPS	200,000	7	2D and 3D	10	[99]
BLVD	Large-scale 5D semantics dataset collected in China's Intelligent Vehicle Proving Center.	Cameras	50,000	7	2D	9	[100]
HEV-I	Dataset includes video clips of real human driving in different intersections in the San Francisco Bay Area.	Cameras, LiDAR	12,000	5	2D and 3D	5	[101]
Ford CVaL	Dataset collected by an autonomous ground vehicle testbed equipped with multiple sensors, collected in Michigan.	Cameras, LiDAR, radar	2000	6	2D and 3D	3	[102]
TbV Dataset	Over 1000 scenarios captured by autonomous vehicles, each log represents a continuous observation of a scene around a self-driving vehicle.	Cameras, LiDAR	10,000	4	2D and 3D	2	[103]
Argoverse 2 Map Change	Temporal annotations indicating map changes within 30 m of an autonomous vehicle.	Cameras, LiDAR, radar	1200	8	2D and 3D	1	[64]
D2CITY	Large-scale collection of dashcam videos collected by vehicles on DiDi's platform.	Cameras, LiDAR	12,000	8	2D and 3D	1	[104]
DADE	Sequences acquired by agents (ego vehicles) within a 5 h time frame, totaling 990k frames.	Cameras, LiDAR	20,000	5	2D and 3D	1	[105]
DMPD	Test set contains images and pedestrian labels captured from a vehicle during a 27 min drive.	Cameras	18,000	4	2D	1	[106]
INDRA	Dataset consisting of 104 videos with annotated road crossing safety labels and vehicle bounding boxes.	Cameras, LiDAR	70,000	5	2D and 3D	1	[107]

Table 2. *Cont.*

Name	Description	Sensors Used	Samples Number	Classes Number	2D/3D	Cited	Ref.
METEOR	Consists of 1000+ one-minute video clips with annotated frames and bounding boxes for surrounding vehicles and traffic agents.	Cameras, LiDAR, IMU	15,000	7	2D and 3D	1	[108]
METU-VIREF	VIRAT dataset for surveillance containing primarily people and vehicles, aligned with videos from the ILSVRC dataset.	Cameras	12,000	5	2D	1	[109]
RoadTextVQA	Video question-answering dataset for in-vehicle conversations.	Cameras	30,000	4	2D	1	[110]
TLV	Real-world datasets based on NuScenes and Waymo for temporal logic.	Cameras	50,000	3	2D	1	[111]
Vehicle-Rear	Dataset for vehicle identification with high-resolution videos, including make, model, color, and license plates.	Cameras	10,000	4	2D	1	[112]
IMO	Contains images, stereo disparity, and vehicle labels with ground truth annotations.	Cameras, LiDAR, GPS	100,000	6	2D and 3D	0	[113]
LISA Vehicle Detection	Dataset for vehicle detection with video sequences captured at different times and varying traffic conditions.	Cameras	10,000	4	2D	0	[114]

Table 3. Available simulators that can be used to train deep neural models for implementation in autonomous vehicles.

Name	Description	Cited	Ref.
CARLA	Simulator for urban driving with 12 semantic classes, bounding boxes, and vehicle measurements.	1128	[54]
AirSim	Simulator for drones, cars, and more, built on Unreal Engine, with support for SIL and HIL.	248	[115]
TORCS	Driving simulator capable of simulating elements of vehicular dynamics.	91	[55]
V2X-SIM	Synthetic collaborative perception dataset for autonomous driving, collected from both roadside and vehicles.	17	[116]
SUMMIT	Supports a wide range of applications, including perception, control, planning, and end-to-end learning.	13	[117]
CVRPTW	Instances of the Capacitated Vehicle Routing Problem with Time Windows for various customer nodes.	7	[118]
CARL	Control suite extended with physics context features for AI training.	6	[119]
3D VTSim	Dataset collected using driving simulation for accurate 3D bounding box annotations.	4	[120]
MUAD	Dataset with realistic synthetic images under diverse weather conditions, annotated for multiple tasks.	3	[121]
SDN	Navigation benchmark with trials and control streams developed to evaluate dialogue moves and physical navigation actions.	2	[122]
MULTIROTOR-GYM	Multirotor gym environment for learning control policies for UAVs.	2	[123,124]
DEAP CITY	City pollution data, including daily pollutant and meteorological features, alongside total vehicle mileage.	1	[125]
EviLOG	Real-world lidar point clouds from a test vehicle with the same LiDAR setup as simulated LiDAR.	1	[126]

3.1. Image Collections

Table 1 enumerates various image datasets crucial for training deep neural models in autonomous driving. The Cityscapes dataset emerges as the most frequently cited

(3411 citations), offering extensive semantic, instance, and pixel-level annotations across diverse classes. Other notable datasets include Waymo (398 citations) and KITTI-360 (181 citations), which offer comprehensive sensor data and advanced 2D/3D annotations, respectively.

Datasets such as IDD and INTERACTION provide insights into road scene comprehension and interactive traffic participant behavior. For specialized tasks, SemanticPOSS focuses on 3D semantic segmentation, while WoodScape offers a unique fisheye camera perspective. Datasets like Lost and Found and ROAD Anomaly are tailored for obstacle detection and anomaly identification.

Several datasets emphasize stereo vision and uncertainty estimation. Fishyscapes, PandaSet, and KITTI Road supply sensor-rich data and road/lane benchmarks, respectively, while Talk2Car facilitates natural language grounding in visuals. Other contributions include event-based data (MVSEC), urban navigation aids (KAIST Urban), and comprehensive 3D annotations (Cityscapes 3D).

Niche datasets such as RailSem19 and EuroCity Persons cater to rail scene understanding and urban pedestrian detection. Emerging datasets like DOLPHINS and PSI explore V2X network testing and pedestrian intent predictions. TICaM and Zenseact focus on vehicle interior monitoring and extensive sensor data across Europe.

Anomaly instance segmentation is addressed by OoDIS, while real-world interaction scenarios are captured in LOOK (three citations). The diversity and thematic grouping of these datasets underscore their relevance and application in developing robust autonomous vehicle systems.

The presented datasets showcase diverse strengths and weaknesses targeted towards various autonomous driving research applications. Data richness in terms of sample size is notable in datasets such as Waymo, KITTI-360, and WoodScape, enabling comprehensive model training and validation. KITTI-360 and Cityscapes offer extensive 2D and 3D semantic annotations, making them robust for developing versatile algorithms. Conversely, subsets like ROAD Anomaly and PSI are limited by their smaller sample sizes, potentially constraining their utility for generalized model training. Annotation depth varies significantly, with WoodScape providing the maximum class diversity (40 classes) compared to more focused datasets like EuroCity Persons and DOLPHINS. Sensor diversity is a prominent advantage seen in several datasets, with combinations of cameras, LiDAR, GNSS/IMU, and stereo cameras (e.g., KAIST Urban and Waymo), providing a step towards achieving comprehensive environmental perception. However, datasets focused on specific conditions or sensors, such as SemanticPOSS (solely LiDAR) and Fishyscapes (cameras only), may limit cross-modal insights. Citation metrics reveal the broader academic acceptance and utility of some datasets, where Cityscapes and KITTI-360 are highly referenced, underscoring their significance and reliability within the research community. Paying balanced attention to sensor variety, annotation detail, and data volume is crucial for advancing autonomous driving technologies.

3.2. Video Stocks

Table 2 lists various video datasets used for training deep neural models in autonomous driving, highlighting the key datasets and their applications. The nuScenes dataset stands out as the most widely cited (1695 citations) with its comprehensive suite of 32-beam LiDAR, six cameras, and radars providing full 360° coverage, crucial for holistic vehicle perception. The Talk2Car dataset is based on the nuScenes dataset. It is the first dataset designed for object referencing with natural language commands for autonomous vehicles. Talk2Car enables the study of how autonomous cars can understand and execute passenger commands in urban settings.

Other significant datasets include Virtual KITTI, with its photo-realistic synthetic videos, and CULane, which focuses on lane detection through camera footage collected in Beijing. The ApolloScape dataset, notable for its extensive collection of over 140,000 video frames, captures diverse locations across China.

For specific driving scenarios, ROAD evaluates the ability of autonomous vehicles to recognize road events through annotated videos. Datasets like V2V4Real encompass multi-modal sensor data from two vehicles driving together. TITAN and BrnoCompSpeed include labeled video clips with odometry and speed measurements, respectively.

Additionally, CCD features real traffic accident videos from dashcams, and BLVD offers a large-scale semantic dataset from China's Intelligent Vehicle Proving Center. The HEV-I dataset includes real human driving clips at San Francisco intersections, while Ford CVaL provides multi-sensor data from Michigan.

More niche datasets, such as the TbV Dataset, with over 1000 autonomous vehicle scenarios, and Argoverse 2 Map Change, indicating temporal map changes, serve specialized purposes. Datasets like D2CITY, DADE, and DMPD each have minimal citations, but contribute with extensive dashcam videos and pedestrian labels.

Emerging and less-cited datasets include INDRA for road-crossing safety, METEOR with annotated traffic clips, METU-VIREF for surveillance, and RoadTextVQA for in-vehicle conversation question answering. TLV, Vehicle-Rear, and IMO offer enriched vehicle identification and traffic data, albeit with very few citations.

Overall, these video datasets collectively support the development of robust autonomous driving models, catering to various aspects such as object detection, lane marking, event recognition, and real-world driving behavior.

The expanded dataset table highlights the varying strengths and weaknesses pertinent to autonomous driving research. Notably, nuScenes excels with its comprehensive 360° sensor suite, featuring multiple sensors (cameras, LiDAR, RADAR, GPS, and IMU) and a substantial volume of samples (1.4 M), making it a powerful resource for multi-modal perception studies, evidenced by its high citation count. On the contrary, datasets such as V2V4Real and Ford CVaL, although equipped with multiple sensors, are limited by smaller sample sizes, constraining any extensive testing capabilities. Virtual KITTI provides a unique edge with its photo-realistic synthetic data, contributing to experimental control in simulation environments, albeit with a relatively moderate citation impact.

Datasets like CULane and ApolloScape, with their sizable sample counts, are advantageous for focused domain tasks, such as lane detection and general driving scene understanding, respectively. However, some datasets such as TITAN and BrnoCompSpeed offer specific yet limited applications due to their restricted sample sizes and fewer annotated classes. The research utility of datasets like CCD and DADE is highlighted by their moderate sample volumes; however, their lack of citations indicates potential underutilization or emerging utility in the field.

Several datasets like HEV-I and V2V4Real are noted for their multi-modal sensor data collected in specific geographical regions or setups, presenting strengths in context-specific model validations. Conversely, newly introduced datasets such as Argoverse 2 Map Change and DADE are just beginning to gain traction, reflected by their low citation counts yet promising utility for future research.

Sensor diversity across datasets fosters the development of robust perception algorithms, while variations in sample size, annotation depth, and class diversity present both opportunities and hindrances. Citation metrics further underscore the importance and acceptance within the community, with datasets like nuScenes and Virtual KITTI leading in having a broader academic influence. Balancing these factors is crucial for advancing the next generation of autonomous driving systems.

3.3. Simulators

Table 3 presents various simulators designed to train deep neural models for autonomous driving. The most widely cited simulator is CARLA (1128 citations), offering versatile urban driving simulations with detailed annotations, including 12 semantic classes, bounding boxes, and vehicle measurements.

AirSim, developed on Unreal Engine, supports simulation-in-the-loop (SIL) and hardware-in-the-loop (HIL) for drones and cars, providing a valuable tool for multi-platform simulations. The TORCS simulator is notable for simulating detailed vehicular dynamics, facilitating advanced driving behavior analyses [127].

V2X-SIM focuses on synthetic collaborative perception, capturing data both from the roadside and vehicles, emphasizing vehicle-to-everything (V2X) communications. SUMMIT is a versatile simulator supporting applications across perception, control, planning, and end-to-end learning.

Other simulators include CVRPTW, which addresses the Capacitated Vehicle Routing Problem with Time Windows, and CARL, which enhances AI training with physics context features. The 3D VTSim provides accurate 3D bounding box annotations through simulation, supporting precise object detection tasks.

The MUAD simulator generates realistic synthetic images under various weather conditions, annotated for multiple tasks. SDN serves as a benchmark for navigation tasks, evaluating the integration of dialogue and physical navigation actions. MULTIROTOR-GYM is dedicated to learning the control policies for UAVs within a gym environment.

Emerging simulators such as DEAP CITY, which includes city pollution data, and EviLOG, with real-world LiDAR point cloud data, extend the utility of simulators to environmental analytics and precise sensor data simulations.

These simulators collectively offer comprehensive tools essential for advancing autonomous driving technologies, focusing on aspects varying from control policies to environmental context and collaborative perception.

This augmented dataset overview brings into focus an interesting array of simulation and synthetic data sources aimed at enhancing autonomy in driving and other vehicular contexts. Leading the way is CARLA, a widely cited urban driving simulator with robust semantic annotations, making it highly valuable for both academic and industry research. Similarly, AirSim offers a versatile simulation platform for both aerial and ground vehicles, supported by its moderate citation count, indicating its cross-domain applicability.

Comparatively, TORCS provides a specialized focus on vehicular dynamics simulations with fewer citations, suggesting a niche but steady usage. V2X-SIM stands out in the emerging field of collaborative perception despite its relatively low citation count, indicating ongoing exploration in synthetic data from both roadside and vehicular perspectives.

Multi-disciplinary simulators like SUMMIT and CARL, although less cited, feature support for comprehensive applications ranging from perception to end-to-end learning, providing potential growth areas for research in autonomous systems. The specific problem-oriented dataset CVRPTW emphasizes complex logistic problems and demonstrates modest citations, emphasizing its specialized use case in routing and planning scenarios.

Datasets like 3D VTSim and MUAD, focusing on accurate 3D bounding boxes and diverse weather condition simulations, respectively, are in the nascent stage of adoption but show promise for detailed perception tasks. Benchmarks like SDN and MULTIROTOR-GYM offer integrated environments for evaluating navigation and control policies, particularly for UAVs, though their citations remain low, suggesting potential for traffic in specific research contexts.

Less cited datasets such as DEAP CITY and EviLOG, with a focus on environmental impact and real-world LiDAR point clouds, respectively, seek to bridge the gap between simulations and practical applications, yet are still gaining traction.

Overall, citation metrics reflect the varying degrees of community engagement, with prominent datasets like CARLA and AirSim demonstrating a significant impact. The diversity in applications, ranging from simulations in dynamic environments to control policy evaluation, underscores the continual expansion of the capabilities and deployment scenarios of autonomous systems, necessitating balanced dataset selection to meet specific research and development needs.

4. Comparison of Selected AI Models

The comparative testing of various models on selected datasets is critical for evaluating and understanding the performances of different algorithms and systems in autonomous vehicles. Such comparisons provide insights into the strengths and weaknesses of various approaches, guiding researchers towards more effective and robust solutions. Comparative tests on key datasets like Cityscapes, nuScenes and CARLA offer insights into model performance across tasks like 3D object detection, semantic segmentation, and driving proficiency in simulation conditions.

4.1. Semantic Segmentation

Semantic segmentation is the process of assigning a label to each pixel in an image so that pixels with the same label have some common characteristics. In the case of the Cityscapes dataset, semantic segmentation involves assigning labels (e.g., "tree", "car", "sidewalk", etc.) to pixels in images of urban scenes (Figure 2). Models glassed using data from Cityscapes generate predictions that can be compared to benchmark data.

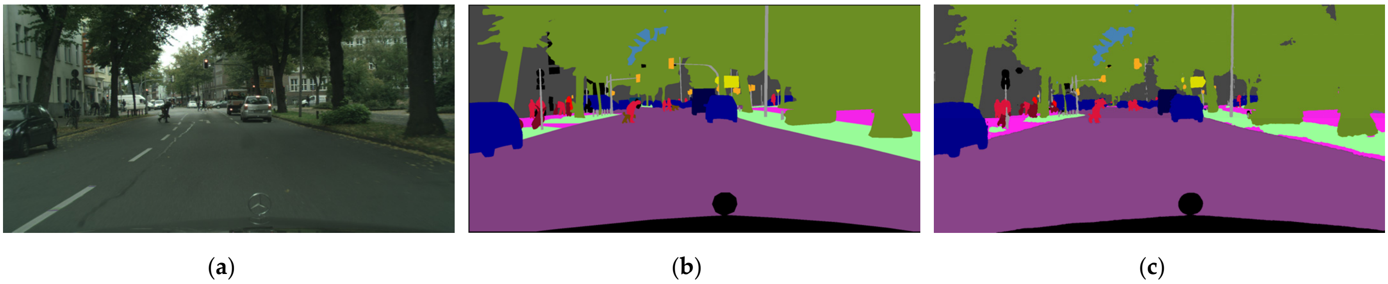

Figure 2. An example of semantic segmentation on the Cityscapes set where: (**a**) original image; (**b**) ground truth image; and (**c**) model prediction result [128].

Table 4 presents a comparative analysis of various semantic segmentation models on the Cityscapes dataset, evaluated in terms of mean Intersection over Union (mIoU) and frames per second (fps) on different GPUs. The analysis highlights both segmentation accuracy and real-time performance.

Table 4. Comparison of selected models capable of semantic segmentation tested on a set of images obtained from Cityscapes.

Model	mIoU	Fps (GPU)	Year	Ref.
YOLOv8-DSAF	96.1%	192 (3090)	2024	[129]
RT-DETR	89.1%	200 (3090)	2023	[130]
YOLOv8n	87.2%	178 (3090)	2023	[131]
VLTSeg	86.4%	76 (n/a)	2023	[132]
PIDNet-L	80.6%	31.1 (3090)	2023	[133]
SFNet-R18	80.4%	25.7 (1080Ti)	2020	[134]
PIDNet-M	79.8%	42.2 (3090)	2022	[133]
PIDNet-S	78.6%	93.2 (3090)	2022	[133]
RegSeg	78.3%	30 (n/a)	2021	[135]
PP-LiteSeg-B2	77.5%	102.6 (1080Ti)	2022	[136]
DDRNet-23-slim	77.4%	101.6 (2080Ti)	2021	[137]
DPNet (DeepLabV3Plus)	76.8%	20 (n/a)	2023	[138]
STDC2-75	76.8%	97.0 (1080Ti)	2021	[139]
U-HarDNet-70	75.9%	53 (1080Ti)	2019	[140]
HyperSeg-M	75.8%	36.9 (n/a)	2020	[141]
SwiftNetRN-18	75.5%	39.9 (n/a)	2019	[142]
STDC1-75	75.3%	126.7 (n/a)	2021	[139]
BiSeNet V2-Large	75.3%	47.3 (n/a)	2021	[143]
TD4-BISE18	74.9%	47.6 (Titan X)	2020	[144]
PP-LiteSeg-T2	74.9%	143.6 (1080Ti)	2022	[136]

Among the models evaluated in Table 4 for their semantic segmentation proficiency on Cityscapes images, YOLOv8-DSAF and RT-DETR stand out prominently. YOLOv8-DSAF achieves the highest mIoU (mean Intersection over Union) of 96.1%, which indicates its superior accuracy in segmenting images. Additionally, it operates at an impressive rate of 192 fps on an NVIDIA 3090 GPU, making it not only highly accurate, but also exceptionally fast. This model uses three key modules: Depthwise Separatable Convolution (DSConv), a Dual-Path Attention Gate (DPAG), and a Feature Enhancement Module (FEM). With these technologies, YOLOv8-DSAF achieves high detection accuracy in dynamic urban scenarios. RT-DETR, although slightly less accurate with an mIoU of 89.1%, surpasses YOLOv8-DSAF in speed, achieving 200 fps on the same GPU, which makes it notable for real-time applications. RT-DETR is the first fully end-to-end real-time object detector that eliminates the need for Non-Maximum Suppression (NMS). Moreover, models like YOLOv8n and VLTSeg also show high mIoU scores of 87.2% and 86.4%, respectively, but they fall behind in terms of fps, with YOLOv8n operating at 178 fps and VLTSeg at a significantly lower 76 fps. These metrics highlight YOLOv8-DSAF and RT-DETR as the most efficient and effective models in terms of both segmentation accuracy and processing speed.

This comparative analysis underscores the progressive enhancement in semantic segmentation capabilities, reflecting improvements in both mIoU and fps over time, driving the efficiency and applicability of these models in real-world scenarios.

4.2. 3D Object Detection

The detection of 3D objects in the nuScenes set involves identifying and locating objects in three-dimensional space based on data from cameras (Figure 3). In this task, algorithms analyze images from cameras and then detect and describe objects such as vehicles, pedestrians, and road signs.

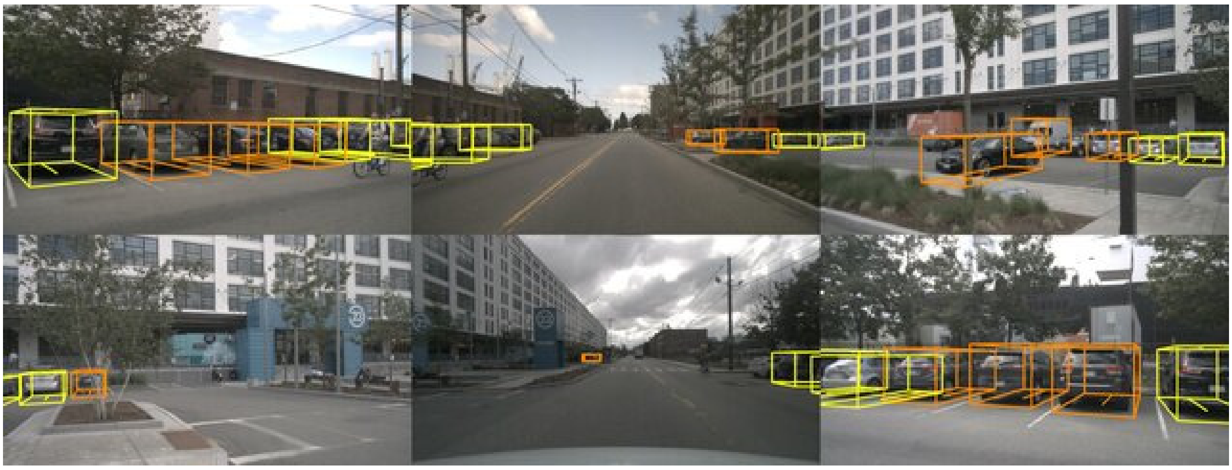

Figure 3. Sample of NuScenes labels. Objects on a single image are colored in orange, while those on two consecutive cameras are shown in yellow [145].

The performances of various 3D object detection models evaluated on the nuScenes dataset are presented in Table 5. The models are compared based on several metrics, as follows: NDS (NuScenes Detection Score), mAP (mean Average Precision), mATE (mean Average Translation Error), mASE (mean Average Scale Error), mAOE (mean Average Orientation Error), mAVE (mean Average Velocity Error), and mAAE (mean Average Attribute Error), along with the year of publication and reference.

The EA-LSS model demonstrates the highest performance, with an NDS of 0.78 and mAP of 0.77, showcasing commendable results across all metrics and introducing an edge-aware framework for enhanced object boundary delineation and spatial accuracy in 3D Bird’s Eye View (BEV) contexts.

Table 5. Comparison of selected models capable of detecting 3D objects tested on the nuScenes video set.

Model	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE	Year	Ref.
EA-LSS	0.78	0.77	0.23	0.21	0.28	0.20	0.12	2023	[146]
BEVFusion-e	0.76	0.75	0.24	0.23	0.32	0.22	0.13	2022	[147]
FocalFormer3D-F	0.75	0.72	0.25	0.24	0.33	0.23	0.13	2023	[148]
UniTR	0.75	0.71	0.24	0.23	0.26	0.24	0.13	2023	[149]
FocalFormer3D-TTA	0.74	0.71	0.24	0.24	0.32	0.20	0.13	2023	[148]
MEFormer	0.74	0.72	0.27	0.24	0.30	0.27	0.11	2024	[150]
3D Dual-Fusion_T	0.73	0.71	0.26	0.24	0.33	0.27	0.13	2022	[151]
FocalFormer3D-L	0.73	0.69	0.25	0.24	0.34	0.22	0.13	2023	[148]
MGTANet	0.73	0.67	0.25	0.23	0.31	0.19	0.12	2022	[152]
CenterPoint	0.71	0.67	0.25	0.24	0.35	0.25	0.14	2020	[153]
SSN	0.62	0.51	0.34	0.24	0.43	0.27	0.09	2020	[154]

Other notable models include BEVFusion, which integrates multi-sensor data into a unified BEV representation to support concurrent tasks such as detection, segmentation, and tracking. FocalFormer3D employs a targeted mechanism to prioritize challenging instances within the 3D space, leveraging focal attention combined with transformer networks for an improved detection accuracy in complex environments. UniTR presents a unified transformer framework for multi-modal BEV representation, efficiently fusing data from LiDAR and cameras to produce comprehensive outputs.

The 3D Dual-Fusion model enhances detection through a dual-query fusion approach, integrating both camera and LiDAR data to leverage the strengths of each sensor type. MGTANet incorporates a Long Short-Term Motion-Guided Temporal Attention mechanism to improve the detection of moving objects by encoding sequential LiDAR points. CenterPoint focuses on center-based 3D detection and tracking to simplify object localization and association across frames. Lastly, SSN uses shape signatures to encode geometric features for robust multi-class object detection from point clouds.

These models collectively contribute to advancements in 3D object detection by addressing a range of challenges through innovative frameworks and mechanisms tailored to enhance precision, recall, and overall detection performance.

4.3. Results on the CARLA Platform

CARLA Leaderboard 1.0 is an earlier version of the platform for assessing autonomous driving algorithms in the CARLA simulator, characterized by photorealistic graphics (Figure 4). Algorithms are evaluated based on simpler scenarios. In turn, CARLA Leaderboard 2.0 is the latest version, introducing more complex scenarios, such as door-opening maneuvers or giving way to emergency vehicles. Leaderboard 2.0 also supports more sensors, including eight RGB cameras, two LIDAR scanners, and four radars.

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

Figure 4. Examples of images from the CARLA simulator: (**a**) street view and (**b**) view from the car's perspective.

Tables 6 and 7 present the performances of the leading deep neural models on the CARLA Leaderboard tasks 1.0 and 2.0, respectively, with metrics including Driving Score, Route Completion, and Infraction Penalty/Score.

Table 6. Performance of leading deep neural models on the CARLA Leaderboard 1.0 task.

Model	Driving Score	Route Completion	Infraction Penalty	Year	Ref.
ReasonNet	79.95	89.89	0.89	2022	[155]
InterFuser	76.18	88.23	0.84	2022	[156]
TGCP	75.14	85.63	0.87	2022	[157]
LAV	61.84	94.46	0.64	2022	[158]
TransFuser	61.18	86.70	0.71	2022	[159]
TransFuser (Reproduced)	55.04	89.65	0.63	2022	[159]
TGCP (Reproduced)	47.91	65.73	0.77	2023	[157]
Latent TransFuser	45.20	66.31	0.72	2022	[159]

Table 7. Performance of leading deep neural models on the CARLA Leaderboard 2.0 task.

Model	Driving Score	Route Completion	Infraction Score	Year	Ref.
CarLLaVA	6.87	18.08	0.42	2024	[160]
CarLLaVA (Map Track)	6.25	18.89	0.39	2024	[160]
TF++ (Map Track)	5.56	11.82	0.47	2024	[161]
TF++	5.18	11.34	0.48	2024	[161]

ReasonNet achieves the highest Driving Score (79.95) and Route Completion (89.89), indicating a superior overall performance. ReasonNet's end-to-end framework integrates temporal and global reasoning to capture dynamic environmental changes and comprehensively understand scene contexts.

LAV demonstrates the highest Route Completion score (94.46) and lowest Infraction Penalty (0.64), highlighting its efficiency in route completion and safety. LAV's approach aggregates data from multiple vehicles to develop robust models capable of handling diverse driving scenarios.

Interpretable Sensor Fusion Transformer integrates data from multiple sensors, offering an improved interpretability and analysis of the fused information.

Trajectory-guided Control Prediction (TCP) utilizes predicted trajectories for control decisions, effectively bridging perception and action planning.

Hidden Biases of End-to-End Driving Models examines the biases in autonomous driving systems, revealing how the training data and model architecture may influence decision making.

TransFuser Models employ transformer-based frameworks for sensor fusion, enhancing decision making through the integration of multi-modal sensor data.

In Table 7, CarLLaVA achieves the highest Driving Score (6.87) and a commendable Route Completion score (18.08), with a low Infraction Score (0.42). It leverages a Vision Language Model for camera-only autonomous driving, integrating visual and linguistic data for advanced decision making.

CarLLaVA (Map Track) excels with the highest Route Completion score (18.89) and lowest Infraction Score (0.39), making it exceptionally robust in route execution and safety.

The TF++ model focuses on uncovering and mitigating biases within autonomous driving systems by integrating bias detection mechanisms with transformer architectures, ensuring more reliable and equitable driving decisions across varied conditions.

This comparative analysis reveals that models like ReasonNet and LAV on CARLA Leaderboard 1.0 demonstrate a high overall performance and route completion efficiency, respectively. For CARLA Leaderboard 2.0, CarLLaVA and its Map Track variation stand out for their advanced route completion and low infraction rates. Advances such as the Interpretable Sensor Fusion Transformer, TCP framework, and initiatives addressing hidden biases highlight the ongoing efforts to enhance the robustness and reliability of autonomous driving systems. These models and methodologies collectively contribute to the progress in achieving more efficient and safer autonomous driving solutions.

4.4. Results on the Talk2Car Platform

The Talk2Car dataset intersects with various research fields, fostering the development of interdisciplinary solutions to enhance the knowledge on grounding natural language in visual space. Annotations were collected with a focus on high-quality, free-form natural language commands to stimulate practical solutions and a realistic task setting, specifically considering autonomous driving where a passenger could control the vehicle's actions through natural language commands. Built on the nuScenes dataset, Talk2Car includes a comprehensive set of sensor modalities such as semantic maps, GPS, LiDAR, RADAR, and 360-degree RGB images annotated with 3D bounding boxes. This diversity of input modalities distinguishes the object referral task in Talk2Car from related challenges that typically lack additional sensor modalities.

Table 8 presents the current leaderboard for the Talk2Car benchmark, showcasing the performances of various models. The evaluation metric used to rank these models is the Intersection over Union (IoU) of the predicted object bounding box and the ground truth bounding box, with a threshold of 0.5. This metric is also known as IoU0.5 or AP50. The table highlights the top-performing models, encouraging continuous improvement and innovation through the submission of new results and models.

Table 8. Talk2Car model rankings.

Model	AP50	Year	Paper
Udeer_HuBo-VLM	76.74	2023	[18]
Deformerable-MDETR	74.4	2021	[162]
FA	73.51	2022	[163]
Stacked VLBert	71.0	2020	[164]

The table summarizes the performances, measured by AP50, of various Vision Language Models tailored for tasks related to autonomous driving and human–robot interaction. The Udder_HuBo-VLM model stands out with the highest AP50 score of 76.74, underscoring its efficacy in handling unified vision–language tasks and marking its relevance among recent advancements from 2023. Following closely, Deformable-MDETR scored 74.4 in 2021, indicating its strength in modulated detection for end-to-end multi-modal systems. The FA model, with an AP50 of 73.51, presented in 2022, exhibits robust capabilities in interpreting and executing commands autonomously. Stacked VLBert achieved 71.0 in 2020, showing a strong performance, though slightly lower than that of newer models. The Vilbert (Base) model, dating back to 2019, yields an AP50 of 68.9, providing foundational work for subsequent advancements, despite having the lowest AP50 among the listed models. Collectively, these models demonstrate a trajectory of improvement over recent years in the integration and performance of vision–language frameworks in autonomous and interactive tasks.

5. Discussion, Limitations, and Future Research Trends

5.1. Discussion

The integration of advanced AI-based image analysis into autonomous vehicles (AVs) has witnessed tremendous progress, driven by sophisticated datasets and cutting-edge AI methodologies. This review outlined several datasets—Cityscapes, NuScenes, CARLA, and Talk2Car—each serving a distinct role in evaluating and training AI models for various AV tasks, including semantic segmentation, 3D object detection, and vehicle control. Advanced

AI models such as Convolutional Neural Networks (CNNs), Transformer models, Generative Adversarial Networks (GANs), and Vision Language Models (VLMs) have been instrumental in pushing the boundaries of autonomous driving technologies.

AI techniques have significantly improved the precision and robustness of AV perception systems. For instance, semantic segmentation models show high mean Intersection over Union (mIoU) scores while maintaining fast processing speeds, evident from the performances of models like YOLOv8-DSAF and RT-DETR on the Cityscapes dataset. In 3D object detection, models such as EA-LSS and BEVFusion leverage data from multiple sensors, enhancing the detection accuracy in complex environments. For vehicle control in virtual environments, frameworks like ReasonNet and CarLLaVA show high driving scores and route completion rates, further solidifying their roles in AV navigation and decision making.

5.2. Limitations

However, despite these advancements, the current state of research reveals several challenges and limitations that need addressing to push the field further.

1. Dataset Diversity and Representation

While datasets like Cityscapes and NuScenes provide extensive imagery and annotations, they predominantly focus on urban environments. This limits the ability of AI models trained on these datasets to generalize across varied geographical and environmental conditions, such as rural roads, adverse weather conditions, and nighttime driving. Datasets that encompass a broader spectrum of driving scenarios and conditions are needed to develop more versatile and robust AV systems.

2. Multi-Sensor Fusion Limitations

Many of the high-performing models rely heavily on multi-sensor data (e.g., cameras, LiDAR, and radar). Ensuring seamless and efficient sensor fusion remains challenging, particularly in terms of data synchronization, calibration, and dealing with inconsistencies across sensor modalities. Further, the cost and complexity of deploying multi-sensor systems can be prohibitive for their widespread adoption.

3. Real-time processing

Achieving high processing speeds without compromising detection and classification accuracy is a challenging trade-off. This becomes critical in real-time applications where latency must be minimized to ensure safety and reliability.

4. Biases in AI models

Models may inadvertently incorporate biases from training datasets, resulting in systemic prejudices that can affect decision making. Therefore, identifying and mitigating these biases remains a significant challenge.

5. Adaptability to Environmental Variability

Autonomous systems must improve their capacity to operate reliably under a wide range of environmental conditions, including adverse weather, poor lighting, and complex urban landscapes. Advanced models need to further refine their ability to dynamically adapt to such conditions.

5.3. Future Research Trends

1. Advanced data collection and annotation

The development of semi-automated and fully automated annotation tools could alleviate the cumbersome data labeling process. Techniques like weak supervision and active learning can substantially reduce human effort while enhancing the quality and diversity of training datasets.

2. Domain adaptation and generalization

Future research could focus on developing models that generalize better across diverse environments. Domain adaptation techniques and unsupervised learning approaches could improve the robustness of AI models, enabling them to perform consistently in varying real-world scenarios.

3. Real-time processing enhancements

Optimizing network architectures to balance accuracy with processing speed will be crucial. Leveraging hardware advancements alongside software optimizations can lead to significant improvements in the real-time application of these models.

4. Mitigating AI biases

Developing techniques to identify, quantify, and mitigate biases within datasets and AI models will be an essential area of future research. Enhancing the transparency and interpretability of AI models can lead to more equitable and trustworthy machine learning systems.

5. Integration of Multimodal Sensory Data

Enhanced sensor fusion techniques that seamlessly integrate data from various sensors (cameras, LiDAR, radar, etc.) are essential for improving contextual understanding and decision-making capabilities. Innovations in multimodal approaches, such as those demonstrated by VLMs, hold significant promise.

6. Simulation and virtual environments

Refining simulation tools like CARLA and integrating them with real-world data can create more effective and versatile training systems for autonomous vehicles. These simulators will be critical for validating algorithms under varied and controlled conditions.

7. Human–machine interaction

Improving the interface between human operators and autonomous systems can enhance safety and trust. Research focusing on intuitive control mechanisms and fail-safe protocols is essential to ensure effective human intervention when necessary.

In summary, while substantial progress has been made in leveraging AI for image analysis in autonomous vehicles, ongoing research and innovation are required to overcome existing limitations and pave the way for more advanced, reliable, and safe autonomous driving systems.

6. Conclusions

Environmental image analysis has emerged as a foundational component in the evolutionary trajectory of autonomous vehicle technology. This review elucidated the multifaceted role that image analysis plays across varying levels of vehicle autonomy, and highlighted the intricate interplay between advanced datasets, cutting-edge AI models, and real-world applications.

The availability and utilization of comprehensive datasets like Cityscapes, nuScenes, and CARLA have revolutionized the training, validation, and benchmarking of autonomous vehicle perception systems. These datasets provide rich and diverse scenarios that are critical for developing robust AI models capable of handling the vast array of driving environments and conditions. The integration of multi-sensor data, including LiDAR, radar, and high-resolution cameras, in these datasets facilitates the development of sophisticated sensor fusion techniques that enhance environmental perception.

The significant improvements in Convolutional Neural Networks (CNNs), Transformer models, Generative Adversarial Networks (GANs), and Vision Language Models (VLMs) have propelled the field of image analysis. Techniques such as EfficientNet and Vision Transformers have demonstrated a remarkable efficiency and accuracy, while advancements in GANs have considerably enriched training datasets and improved their robustness.

The incorporation of federated learning and blockchain technology ensures privacy and security by enabling decentralized model training. This innovative approach allows for continuous learning from diverse data sources without compromising sensitive data.

Models like YOLOv8-DSAF and RT-DETR have set new benchmarks in semantic segmentation with a superior speed and accuracy, crucial for real-time processing. In 3D object detection, models such as EA-LSS and BEVFusion stand out for their precision in identifying and localizing objects in dynamic traffic scenarios.

The assessment of leading models on simulation platforms like CARLA has provided valuable insights into their practical capabilities. For instance, models like ReasonNet and LAV in CARLA Leaderboard 1.0 and CarLLaVA in Leaderboard 2.0 exemplify notable advancements in route completion, driving scores, and infraction minimization. Moreover, the Talk2Car platform emphasizes the growing importance of integrating natural language processing with vision systems, as evidenced by high-performing Vision Language Models like Udder_HuBo-VLM.

The continuous evolution of image analysis technologies is integral to advancing the autonomy of vehicles. As AI models become more sophisticated and datasets more comprehensive, the capabilities of autonomous vehicles will continue to expand, pushing the boundaries of safety, efficiency, and intelligence in transportation. Ongoing research and development, coupled with rigorous testing and validation, will drive the deployment of reliable and sophisticated autonomous systems, ultimately transforming the landscape of modern transportation.

Author Contributions: Conceptualization, M.K.; methodology, M.K. and S.R.; formal analysis, M.K. and S.R.; investigation, M.K. and S.R.; resources, M.K. and S.R.; writing—original draft preparation, M.K. and S.R.; writing—review and editing, M.K. and S.R.; visualization, M.K. and S.R.; supervision, S.W.; project administration, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Shin, S.; Cho, Y.; Lee, S.; Park, J. Assessing Traffic-Flow Safety at Various Levels of Autonomous-Vehicle Market Penetration. *Appl. Sci.* **2024**, *14*, 5453. [CrossRef]
2. Schrader, M.; Hainen, A.; Bittle, J. Extracting Vehicle Trajectories from Partially Overlapping Roadside Radar. *Sensors* **2024**, *24*, 4640. [CrossRef] [PubMed]
3. Booth, L.; Karl, C.; Farrar, V.; Pettigrew, S. Assessing the Impacts of Autonomous Vehicles on Urban Sprawl. *Sustainability* **2024**, *16*, 5551. [CrossRef]
4. Muhovič, J.; Perš, J. Correcting Decalibration of Stereo Cameras in Self-Driving Vehicles. *Sensors* **2020**, *20*, 3241. [CrossRef]
5. Huang, P.; Tian, S.; Su, Y.; Tan, W.; Dong, Y.; Xu, W. IA-CIOU: An Improved IOU Bounding Box Loss Function for SAR Ship Target Detection Methods. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10569–10582. [CrossRef]
6. Lin, Y.H.; Chen, S.Y. Development of an Image Processing Module for Autonomous Underwater Vehicles through Integration of Object Recognition with Stereoscopic Image Reconstruction. In Proceedings of the ASME 2019 38th International Conference on Ocean, Offshore and Arctic Engineering, Glasgow, UK, 9–14 June 2019. [CrossRef]
7. Nian, R.; Liu, F.; He, B. An Early Underwater Artificial Vision Model in Ocean Investigations via Independent Component Analysis. *Sensors* **2013**, *13*, 9104–9131. [CrossRef]
8. He, B.; Zhang, H.; Li, C.; Zhang, S.; Liang, Y.; Yan, T. Autonomous Navigation for Autonomous Underwater Vehicles Based on Information Filters and Active Sensing. *Sensors* **2011**, *11*, 10958–10980. [CrossRef]
9. Kim, J.; Cho, J. Rgdinet: Efficient Onboard Object Detection with Faster r-Cnn for Air-to-Ground Surveillance. *Sensors* **2021**, *21*, 1677. [CrossRef]
10. Salles, R.N.; de Campos Velho, H.F.; Shiguemori, E.H. Automatic Position Estimation Based on Lidar × Lidar Data for Autonomous Aerial Navigation in the Amazon Forest Region. *Remote Sens.* **2022**, *14*, 361. [CrossRef]
11. Yang, T.; Ren, Q.; Zhang, F.; Xie, B.; Ren, H.; Li, J.; Zhang, Y. Hybrid Camera Array-Based UAV Auto-Landing on Moving UGV in GPS-Denied Environment. *Remote Sens.* **2018**, *10*, 1829. [CrossRef]
12. Wang, H.; Lu, E.; Zhao, X.; Xue, J. Vibration and Image Texture Data Fusion-Based Terrain Classification Using WKNN for Tracked Robots. *World Electr. Veh. J.* **2023**, *14*, 214. [CrossRef]

13. Cabezas-Olivenza, M.; Zulueta, E.; Sánchez-Chica, A.; Teso-Fz-betoño, A.; Fernandez-Gamiz, U. Dynamical Analysis of a Navigation Algorithm. *Mathematics* **2021**, *9*, 3139. [CrossRef]
14. Ci, W.; Huang, Y. A Robust Method for Ego-Motion Estimation in Urban Environment Using Stereo Camera. *Sensors* **2016**, *16*, 1704. [CrossRef] [PubMed]
15. Kim, B.J.; Lee, S.B. A Study on the Evaluation Method of Autonomous Emergency Vehicle Braking for Pedestrians Test Using Monocular Cameras. *Appl. Sci.* **2020**, *10*, 4683. [CrossRef]
16. Kim, Y.-W.; Byun, Y.-C.; Krishna, A.V. Portrait Segmentation Using Ensemble of Heterogeneous Deep-Learning Models. *Entropy* **2021**, *23*, 197. [CrossRef]
17. Kim, J. Detection of Road Images Containing a Counterlight Using Multilevel Analysis. *Symmetry* **2021**, *13*, 2210. [CrossRef]
18. Dong, Z.; Zhang, W.; Huang, X.; Ji, H.; Zhan, X.; Chen, J. HuBo-VLM: Unified Vision-Language Model Designed for HUman roBOt Interaction Tasks. *arXiv* **2023**, arXiv:2308.12537.
19. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with Graph Visual Question Answering. *arXiv* **2023**, arXiv:2312.14150.
20. Wang, Y.; Su, X.; Chen, Q.; Zhang, X.; Xi, T.; Yao, K.; Ding, E.; Zhang, G.; Wang, J. OVLW-DETR: Open-Vocabulary Light-Weighted Detection Transformer. *arXiv* **2024**, arXiv:2407.10655.
21. Camenisch, J.; Drijvers, M.; Hanke, T.; Pignolet, Y.-A.; Shoup, V.; Williams, D. Internet Computer Consensus. In Proceedings of the 2022 ACM Symposium on Principles of Distributed Computing; Association for Computing Machinery: New York, NY, USA, 2022; pp. 81–91.
22. Zhu, X.; Li, H.; Yu, Y. Blockchain-Based Privacy Preserving Deep Learning. In Proceedings of the Information Security and Cryptology; Guo, F., Huang, X., Yung, M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 370–383.
23. Shafay, M.; Ahmad, R.W.; Salah, K.; Yaqoob, I.; Jayaraman, R.; Omar, M. Blockchain for Deep Learning: Review and Open Challenges. *Clust. Comput.* **2023**, *26*, 197–221. [CrossRef]
24. International, S. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. *SAE Int.* **2018**, *4970*, 1–5.
25. Wang, Y.F. Computer Vision Analysis for Vehicular Safety Applications. In Proceedings of the International Telemetering Conference, International Foundation for Telemetering, Las Vegas, NV, USA, 1 January 2015; Volume 82, pp. 944–953.
26. Yebes, J.J.; Bergasa, L.M.; García-Garrido, M.Á. Visual Object Recognition with 3D-Aware Features in KITTI Urban Scenes. *Sensors* **2015**, *15*, 9228–9250. [CrossRef] [PubMed]
27. Borhanifar, H.; Jani, H.; Gohari, M.M.; Heydarian, A.H.; Lashkari, M.; Lashkari, M.R. Fast Controling Autonomous Vehicle Based on Real Time Image Processing. In Proceedings of the 2021 International Conference on Field-Programmable Technology (ICFPT), IEEE, Tokyo, Japan, 6–9 December 2021; pp. 1–4.
28. Kumawat, K.; Jain, A.; Tiwari, N. Relevance of Automatic Number Plate Recognition Systems in Vehicle Theft Detection†. *Eng. Proc.* **2023**, *59*, 185. [CrossRef]
29. Lee, S.H.; Lee, S.H. U-Net-Based Learning Using Enhanced Lane Detection with Directional Lane Attention Maps for Various Driving Environments. *Mathematics* **2024**, *12*, 1206. [CrossRef]
30. Somawirata, I.K.; Widodo, K.A.; Utaminingrum, F.; Achmadi, S. Road Detection Based on Region Grid Analysis Using Structural Similarity. In Proceedings of the 2020 IEEE 4th International Conference on Frontiers of Sensors Technologies (ICFST), IEEE, Beijing, China, 6 November 2020; pp. 63–66.
31. Kaladevi, R.; Shanmugasundaram, H.; Karthikeyan, R. Lane Detection Using Deep Learning Approach. In Proceedings of the 2022 1st International Conference on Computational Science and Technology (ICCST), IEEE, Chennai, India, 9–10 November 2022; pp. 945–949.
32. Navarro, P.J.; Miller, L.; Rosique, F.; Fernández-Isla, C.; Gila-Navarro, A. End-to-End Deep Neural Network Architectures for Speed and Steering Wheel Angle Prediction in Autonomous Driving. *Electronics* **2021**, *10*, 1266. [CrossRef]
33. Itu, R.; Danescu, R. Fully Convolutional Neural Network for Vehicle Speed and Emergency-Brake Prediction. *Sensors* **2024**, *24*, 212. [CrossRef]
34. Hu, C.; Wang, H. Enhancing Rainy Weather Driving: Deep Unfolding Network with PGD Algorithm for Single Image Deraining. *IEEE Access* **2023**, *11*, 57616–57626. [CrossRef]
35. Saravanarajan, V.S.; Chen, R.C.; Hsieh, C.H.; Chen, L.S. Improving Semantic Segmentation Under Hazy Weather for Autonomous Vehicles Using Explainable Artificial Intelligence and Adaptive Dehazing Approach. *IEEE Access* **2023**, *11*, 38194–38207. [CrossRef]
36. 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. [CrossRef]
37. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27 June–1 July 2016.
38. Yao, C.; Liu, X.; Wang, J.; Cheng, Y. Optimized Design of EdgeBoard Intelligent Vehicle Based on PP-YOLOE+. *Sensors* **2024**, *24*, 3180. [CrossRef]
39. Strzelecki, M.H.; Strąkowska, M.; Kozłowski, M.; Urbańczyk, T.; Wielowieyska-Szybińska, D.; Kociołek, M. Skin Lesion Detection Algorithms in Whole Body Images. *Sensors* **2021**, *21*, 6639. [CrossRef] [PubMed]
40. Mahaur, B.; Mishra, K. Small-Object Detection Based on YOLOv5 in Autonomous Driving Systems. *Pattern Recognit. Lett.* **2023**, *168*, 115–122. [CrossRef]

41. Wang, H.; Liu, C.; Cai, Y.; Chen, L.; Li, Y. YOLOv8-QSD: An Improved Small Object Detection Algorithm for Autonomous Vehicles Based on YOLOv8. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2513916. [CrossRef]
42. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
43. Feldsar, B.; Mayer, R.; Rauber, A. Detecting Adversarial Examples Using Surrogate Models. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1796–1825. [CrossRef]
44. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
45. Hu, S.; Liu, J.; Kang, Z. DeepLabV3+/Efficientnet Hybrid Network-Based Scene Area Judgment for the Mars Unmanned Vehicle System. *Sensors* **2021**, *21*, 8136. [CrossRef]
46. Zhai, X.; Kolesnikov, A.; Houlsby, N.; Beyer, L. Scaling Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022.
47. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [CrossRef]
48. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9. [CrossRef]
49. Fu, Y.; Jain, A.; Di, X.; Chen, X.; Mo, Z. DriveGenVLM: Real-World Video Generation for Vision Language Model Based Autonomous Driving. *arXiv* **2024**, arXiv:2408.16647.
50. Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B.L.; Zimmer, W.; Cao, H.; Knoll, A.C. Vision Language Models in Autonomous Driving: A Survey and Outlook. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2024.
51. Gopalkrishnan, A.; Greer, R.; Trivedi, M. Multi-Frame, Lightweight & Efficient Vision-Language Models for Question Answering in Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2024.
52. Liao, Y.; Xie, J.; Geiger, A. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3292–3310. [CrossRef]
53. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**, arXiv:1412.6572.
54. Nikolenko, S.I. Synthetic Data for Deep Learning. In *Springer Optimization and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2019.
55. Santara, A.; Rudra, S.; Buridi, S.A.; Kaushik, M.; Naik, A.; Kaul, B.; Ravindran, B. Madras: Multi Agent Driving Simulator. *J. Artif. Intell. Res.* **2021**, *70*, 1517–1555. [CrossRef]
56. Zheng, K.; Wei, M.; Sun, G.; Anas, B.; Li, Y. Using Vehicle Synthesis Generative Adversarial Networks to Improve Vehicle Detection in Remote Sensing Images. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 390. [CrossRef]
57. Shatnawi, M.; Bani Younes, M. An Enhanced Model for Detecting and Classifying Emergency Vehicles Using a Generative Adversarial Network (GAN). *Vehicles* **2024**, *6*, 1114–1139. [CrossRef]
58. Chen, Z.; Zhang, J.; Zhang, Y.; Huang, Z. Traffic Accident Data Generation Based on Improved Generative Adversarial Networks. *Sensors* **2021**, *21*, 5767. [CrossRef]
59. Zhou, Y.; Fu, R.; Wang, C.; Zhang, R. Modeling Car-Following Behaviors and Driving Styles with Generative Adversarial Imitation Learning. *Sensors* **2020**, *20*, 5034. [CrossRef]
60. Musunuri, Y.R.; Kwon, O.-S.; Kung, S.-Y. SRODNet: Object Detection Network Based on Super Resolution for Autonomous Vehicles. *Remote Sens.* **2022**, *14*, 6270. [CrossRef]
61. Choi, W.; Heo, J.; Ahn, C. Development of Road Surface Detection Algorithm Using Cyclegan-Augmented Dataset. *Sensors* **2021**, *21*, 7769. [CrossRef]
62. Lee, D. Driving Safety Area Classification for Automated Vehicles Based on Data Augmentation Using Generative Models. *Sustainability* **2024**, *16*, 4337. [CrossRef]
63. Sighencea, B.I.; Stanciu, R.I.; Căleanu, C.D. A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction. *Sensors* **2021**, *21*, 7543. [CrossRef]
64. Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J.K.; et al. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks). *arXiv* **2021**, arXiv:2301.00493v1.
65. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27 June–1 July 2016; pp. 3213–3223.
66. Waymo—Self-Driving Cars—Autonomous Vehicles—Ride-Hail. Available online: https://waymo.com/ (accessed on 17 July 2024).
67. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]

68. Varma, G.; Subramanian, A.; Namboodiri, A.; Chandraker, M.; Jawahar, C. IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Santa Rosa, CA, USA, 7–11 January 2019; pp. 1743–1751.
69. Zhan, W.; Sun, L.; Wang, D.; Shi, H.; Clausse, A.; Naumann, M.; Kümmerle, J.; Königshof, H.; Stiller, C.; de La Fortelle, A.; et al. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv* **2019**, arXiv:1910.03088.
70. Pan, Y.; Gao, B.; Mei, J.; Geng, S.; Li, C.; Zhao, H. Semanticposs: A Point Cloud Dataset with Large Quantity of Dynamic Instances. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), IEEE, Las Vegas, NV, USA, 19–23 September 2020; pp. 687–693.
71. Yogamani, S.; Hughes, C.; Horgan, J.; Sistu, G.; Varley, P.; O'Dea, D.; Uricár, M.; Milz, S.; Simon, M.; Amende, K.; et al. WoodScape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving. *arXiv* **2019**, arXiv:1905.01489.
72. Pinggera, P.; Ramos, S.; Gehrig, S.; Franke, U.; Rother, C.; Mester, R. Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles. In Proceedings of the RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, South Korea, 9–14 October 2016; pp. 1099–1106.
73. Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; Zhou, B. DrivingStereo: A Large-Scale Dataset for Stereo Matching in Autonomous Driving Scenarios. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
74. Blum, H.; Sarlin, P.-E.; Nieto, J.; Siegwart, R.; Cadena, C. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3119–3135. [CrossRef]
75. Lis, K.; Nakka, K.K.; Fua, P.; Salzmann, M. Detecting the Unexpected via Image Resynthesis. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 27 October–2 November 2019; pp. 2152–2161.
76. Xiao, P.; Shao, Z.; Hao, S.; Zhang, Z.; Chai, X.; Jiao, J.; Li, Z.; Wu, J.; Sun, K.; Jiang, K.; et al. Pandaset: Advanced Sensor Suite Dataset for Autonomous Driving. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE, Indianapolis, IN, USA, 19–22 September 2021; pp. 3095–3101.
77. Fritsch, J.; Kuehnl, T.; Geiger, A. A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms. In Proceedings of the International Conference on Intelligent Transportation Systems (ITSC), The Hague, The Netherlands, 22–25 September 2013.
78. Zhu, A.Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; Daniilidis, K. The Multivehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2032–2039. [CrossRef]
79. Jeong, J.; Cho, Y.; Shin, Y.-S.; Roh, H.; Kim, A. Complex Urban Dataset with Multi-Level Sensors from Highly Diverse Urban Environments. *Int. J. Robot. Res.* **2019**, *38*, 642–657. [CrossRef]
80. Zendel, O.; Schörghuber, M.; Rainer, B.; Murschitz, M.; Beleznai, C. Unifying Panoptic Segmentation for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 21351–21360.
81. Chan, R.; Lis, K.; Uhlemeyer, S.; Blum, H.; Honari, S.; Siegwart, R.; Fua, P.; Salzmann, M.; Rottmann, M. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. *arXiv* **2021**, arXiv:2104.14812v2.
82. Braun, M.; Krebs, S.; Flohr, F.B.; Gavrila, D.M. EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1844–1861. [CrossRef]
83. Mao, R.; Guo, J.; Jia, Y.; Sun, Y.; Zhou, S.; Niu, Z. DOLPHINS: Dataset for Collaborative Perception Enabled Harmonious and Interconnected Self-Driving. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 1–4 December 2022; pp. 4361–4377.
84. Chen, T.; Jing, T.; Tian, R.; Chen, Y.; Domeyer, J.; Toyoda, H.; Sherony, R.; Ding, Z. Psi: A Pedestrian Behavior Dataset for Socially Intelligent Autonomous Car. *arXiv* **2021**, arXiv:2112.02604.
85. Jing, T.; Xia, H.; Tian, R.; Ding, H.; Luo, X.; Domeyer, J.; Sherony, R.; Ding, Z. Inaction: Interpretable Action Decision Making for Autonomous Driving. In Proceedings of the European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2022; pp. 370–387.
86. Katrolia, J.S.; El-Sherif, A.; Feld, H.; Mirbach, B.; Rambach, J.R.; Stricker, D. TICaM: A Time-of-Flight In-Car Cabin Monitoring Dataset. In Proceedings of the 32nd British Machine Vision Conference 2021, BMVC 2021, Online, 22–25 November 2021; BMVA Press: Oxford, UK, 2021; p. 277.
87. Alibeigi, M.; Ljungbergh, W.; Tonderski, A.; Hess, G.; Lilja, A.; Lindström, C.; Motorniuk, D.; Fu, J.; Widahl, J.; Petersson, C. Zenseact Open Dataset: A Large-Scale and Diverse Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Melbourne, Australia, 2–6 October 2023; pp. 20178–20188.
88. Nekrasov, A.; Zhou, R.; Ackermann, M.; Hermans, A.; Leibe, B.; Rottmann, M. OoDIS: Anomaly Instance Segmentation Benchmark. *arXiv* **2024**, arXiv:2406.11835.
89. Belkada, Y.; Bertoni, L.; Caristan, R.; Mordan, T.; Alahi, A. Do Pedestrians Pay Attention? Eye Contact Detection in the Wild. *arXiv* **2021**, arXiv:2112.04212.
90. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the CVPR, Seattle, WA, USA, 14–19 June 2020.

91. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In Proceedings of the CVPR, Las Vegas, NV, USA, 27 June–1 July 2016.
92. Deruyttere, T.; Vandenhende, S.; Grujicic, D.; Van Gool, L.; Moens, M.-F. Talk2Car: Taking Control of Your Self-Driving Car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Pittsburgh, PA, USA, 2019.
93. Pan, X.; Shi, J.; Luo, P.; Wang, X.; Tang, X. Spatial as Deep: Spatial CNN for Traffic Scene Understanding. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.
94. Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A.S.; Hauswald, L.; Pham, V.H.; Mühlegg, M.; Dorn, S.; et al. A2D2: Audi Autonomous Driving Dataset. *arXiv* **2020**, arXiv:2004.06320.
95. Singh, G.; Akrigg, S.; Di Maio, M.; Fontana, V.; Alitappeh, R.J.; Saha, S.; Jeddisaravi, K.; Yousefi, F.; Culley, J.; Nicholson, T.; et al. ROAD: The ROad Event Awareness Dataset for Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1036–1054. [CrossRef] [PubMed]
96. Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; et al. V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Perception. In Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
97. Malla, S.; Dariush, B.; Choi, C. TITAN: Future Forecast Using Action Priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11186–11196.
98. Sochor, J.; Juránek, R.; Špaňhel, J.; Maršík, L.; Širokỳ, A.; Herout, A.; Zemčík, P. Comprehensive Data Set for Automatic Single Camera Visual Speed Measurement. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 1633–1643. [CrossRef]
99. Bao, W.; Yu, Q.; Kong, Y. Uncertainty-Based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In Proceedings of the ACM Multimedia Conference, Seattle, WA, USA, 12–16 October 2020.
100. Xue, J.; Fang, J.; Li, T.; Zhang, B.; Zhang, P.; Ye, Z.; Dou, J. BLVD: Building A Large-Scale 5D Semantics Benchmark for Autonomous Driving. In Proceedings of the International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019.
101. Yao, Y.; Xu, M.; Choi, C.; Crandall, D.J.; Atkins, E.M.; Dariush, B. Egocentric Vision-Based Future Vehicle Localization for Intelligent Driving Assistance Systems. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), IEEE, Montreal, QC, Canada, 20–24 May 2019; pp. 9711–9717.
102. Pandey, G.; McBride, J.R.; Eustice, R.M. Ford Campus Vision and Lidar Data Set. *Int. J. Robot. Res.* **2011**, *30*, 1543–1552. [CrossRef]
103. Lambert, J.; Hays, J. Trust, but Verify: Cross-Modality Fusion for HD Map Change Detection. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021), online, 6–14 December 2021.
104. Che, Z.; Li, G.; Li, T.; Jiang, B.; Shi, X.; Zhang, X.; Lu, Y.; Wu, G.; Liu, Y.; Ye, J. D ˆ 2-City: A Large-Scale Dashcam Video Dataset of Diverse Traffic Scenarios. *arXiv* **2019**, arXiv:1904.01975.
105. Gérin, B.; Halin, A.; Cioppa, A.; Henry, M.; Ghanem, B.; Macq, B.; De Vleeschouwer, C.; Van Droogenbroeck, M. Multi-Stream Cellular Test-Time Adaptation of Real-Time Models Evolving in Dynamic Environments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 16–21 June 2024; pp. 4472–4482.
106. Yin, G.; Liu, B.; Zhu, H.; Gong, T.; Yu, N. A Large Scale Urban Surveillance Video Dataset for Multiple-Object Tracking and Behavior Analysis. *arXiv* **2019**, arXiv:1904.11784.
107. Brahmbhatt, S. A Dataset and Model for Crossing Indian Roads. In Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing, Bhubaneswar, India, 15–17 December 2022; pp. 1–8.
108. Chandra, R.; Mahajan, M.; Kala, R.; Palugulla, R.; Naidu, C.; Jain, A.; Manocha, D. METEOR: A Massive Dense & Heterogeneous Behavior Dataset for Autonomous Driving. *arXiv* **2021**, arXiv:2109.07648.
109. Anayurt, H.; Ozyegin, S.A.; Cetin, U.; Aktas, U.; Kalkan, S. Searching for Ambiguous Objects in Videos Using Relational Referring Expressions. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019.
110. Tom, G.; Mathew, M.; Garcia-Bordils, S.; Karatzas, D.; Jawahar, C. Reading Between the Lanes: Text VideoQA on the Road. In Proceedings of the International Conference on Document Analysis and Recognition; Springer: Berlin/Heidelberg, Germany, 2023; pp. 137–154.
111. Choi, M.; Goel, H.; Omama, M.; Yang, Y.; Shah, S.; Chinchali, S. Towards Neuro-Symbolic Video Understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 9–13 September 2024.
112. De Oliveira, I.O.; Laroca, R.; Menotti, D.; Fonseca, K.V.O.; Minetto, R. Vehicle-Rear: A New Dataset to Explore Feature Fusion for Vehicle Identification Using Convolutional Neural Networks. *IEEE Access* **2021**, *9*, 101065–101077. [CrossRef]
113. Persson, M.; Forssén, P.-E. Independently Moving Object Trajectories from Sequential Hierarchical Ransac. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP'21); Scitepress Digital Library: Lisbon, Portugal, 2021.
114. Sivaraman, S.; Trivedi, M.M. A General Active-Learning Framework for On-Road Vehicle Recognition and Tracking. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 267–276. [CrossRef]
115. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. Airsim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In Proceedings of the Field and Service Robotics: Results of the 11th International Conference; Springer: Berlin/Heidelberg, Germany, 2018; pp. 621–635.

116. Li, Y.; Ma, D.; An, Z.; Wang, Z.; Zhong, Y.; Chen, S.; Feng, C. V2X-Sim: Multi-Agent Collaborative Perception Dataset and Benchmark for Autonomous Driving. *IEEE Robot. Autom. Lett.* **2022**, *7*, 10914–10921. [CrossRef]
117. Cai, P.; Lee, Y.; Luo, Y.; Hsu, D. SUMMIT: A Simulator for Urban Driving in Massive Mixed Traffic. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, online, 31 May–5 June 2020; pp. 4023–4029.
118. Falkner, J.K.; Schmidt-Thieme, L. Learning to Solve Vehicle Routing Problems with Time Windows through Joint Attention. *arXiv* **2020**, arXiv:2006.09100.
119. Benjamins, C.; Eimer, T.; Schubert, F.; Mohan, A.; Döhler, S.; Biedenkapp, A.; Rosenhahn, B.; Hutter, F.; Lindauer, M. Contextualize Me—The Case for Context in Reinforcement Learning. *arXiv* **2022**, arXiv:2202.04500.
120. Hu, H.-N.; Yang, Y.-H.; Fischer, T.; Darrell, T.; Yu, F.; Sun, M. Monocular Quasi-Dense 3d Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1992–2008. [CrossRef]
121. Franchi, G.; Yu, X.; Bursuc, A.; Tena, A.; Kazmierczak, R.; Dubuisson, S.; Aldea, E.; Filliat, D. MUAD: Multiple Uncertainties for Autonomous Driving, a Benchmark for Multiple Uncertainty Types and Tasks. In Proceedings of the 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, 21–24 November 2022; BMVA Press: Oxford, UK, 2022.
122. Ma, Z.; VanDerPloeg, B.; Bara, C.-P.; Huang, Y.; Kim, E.-I.; Gervits, F.; Marge, M.; Chai, J. DOROTHIE: Spoken Dialogue for Handling Unexpected Situations in Interactive Autonomous Driving Agents. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 4800–4822.
123. Deshpande, A.M.; Kumar, R.; Minai, A.A.; Kumar, M. Developmental Reinforcement Learning of Control Policy of a Quadcopter UAV with Thrust Vectoring Rotors. In Proceedings of the Dynamic Systems and Control Conference; American Society of Mechanical Engineers: New York, NY, USA, 2020; Volume 84287, p. V002T36A011.
124. Deshpande, A.M.; Minai, A.A.; Kumar, M. Robust Deep Reinforcement Learning for Quadcopter Control. *IFAC-Pap.* **2021**, *54*, 90–95. [CrossRef]
125. Bhattacharyya, M.; Nag, S.; Ghosh, U. Deciphering Environmental Air Pollution with Large Scale City Data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*; International Joint Conferences on Artificial Intelligence Organization: Vancouver, BC, Canada, 2022.
126. van Kempen, R.; Lampe, B.; Woopen, T.; Eckstein, L. A Simulation-Based End-to-End Learning Framework for Evidential Occupancy Grid Mapping. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Indianapolis, IN, USA, 19–22 September 2021; pp. 934–939.
127. Rosique, F.; Navarro, P.J.; Fernández, C.; Padilla, A. A Systematic Review of Perception System and Simulators for Autonomous Vehicles Research. *Sensors* **2019**, *19*, 648. [CrossRef]
128. Massimiliano, V. Semantic Segmentation on Cityscapes Using Segmentation Models Pytorch. Available online: https://github.com/massimilianoviola/semantic-segmentation-cityscapes?tab=readme-ov-file (accessed on 8 September 2024).
129. Li, Y.; Huang, Y.; Tao, Q. Improving Real-Time Object Detection in Internet-of-Things Smart City Traffic with YOLOv8-DSAF Method. *Sci. Rep.* **2024**, *14*, 17235. [CrossRef]
130. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs Beat Yolos on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 16–21 June 2024; pp. 16965–16974.
131. Du, Y.; Liu, X.; Yi, Y.; Wei, K. Optimizing Road Safety: Advancements in Lightweight YOLOv8 Models and GhostC2f Design for Real-Time Distracted Driving Detection. *Sensors* **2023**, *23*, 8844. [CrossRef]
132. Hümmer, C.; Schwonberg, M.; Zhong, L.; Cao, H.; Knoll, A.; Gottschalk, H. VLTSeg: Simple Transfer of CLIP-Based Vision-Language Representations for Domain Generalized Semantic Segmentation. *arXiv* **2023**, arXiv:2312.02021.
133. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 19529–19539.
134. Li, X.; You, A.; Zhu, Z.; Zhao, H.; Yang, M.; Yang, K.; Tan, S.; Tong, Y. Semantic Flow for Fast and Accurate Scene Parsing. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 775–793.
135. Gao, R. Rethinking Dilated Convolution for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 4675–4684.
136. Peng, J.; Liu, Y.; Tang, S.; Hao, Y.; Chu, L.; Chen, G.; Wu, Z.; Chen, Z.; Yu, Z.; Du, Y.; et al. Pp-Liteseg: A Superior Real-Time Semantic Segmentation Model. *arXiv* **2022**, arXiv:2204.02681.
137. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Road Scenes. *arXiv* **2021**, arXiv:2101.06085.
138. Wang, J.; Zhang, X.; Yan, T.; Tan, A. Dpnet: Dual-Pyramid Semantic Segmentation Network Based on Improved Deeplabv3 Plus. *Electronics* **2023**, *12*, 3161. [CrossRef]
139. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking Bisenet for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, online, 19–25 June 2021; pp. 9716–9725.

140. Chao, P.; Kao, C.-Y.; Ruan, Y.-S.; Huang, C.-H.; Lin, Y.-L. Hardnet: A Low Memory Traffic Network. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 3552–3561.
141. Nirkin, Y.; Wolf, L.; Hassner, T. Hyperseg: Patch-Wise Hypernetwork for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, online, 19–25 June 2021; pp. 4061–4070.
142. Orsic, M.; Kreso, I.; Bevandic, P.; Segvic, S. In Defense of Pre-Trained Imagenet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12607–12616.
143. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
144. Hu, P.; Caba, F.; Wang, O.; Lin, Z.; Sclaroff, S.; Perazzi, F. Temporally Distributed Networks for Fast Video Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, online, 14–19 June 2020; pp. 8818–8827.
145. Cortés, I.; Beltrán, J.; de la Escalera, A.; García, F. siaNMS: Non-Maximum Suppression with Siamese Networks for Multi-Camera 3D Object Detection. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), IEEE, online, 19–23 September 2020; pp. 933–938.
146. Hu, H.; Wang, F.; Su, J.; Wang, Y.; Hu, L.; Fang, W.; Xu, J.; Zhang, Z. Ea-Lss: Edge-Aware Lift-Splat-Shot Framework for 3d Bev Object Detection. *arXiv* **2023**, arXiv:2303.17895.
147. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. Bevfusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, London, UK, 29 May–2 June 2023; pp. 2774–2781.
148. Chen, Y.; Yu, Z.; Chen, Y.; Lan, S.; Anandkumar, A.; Jia, J.; Alvarez, J.M. Focalformer3d: Focusing on Hard Instance for 3d Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Melbourne, Australia, 2–6 October 2023; pp. 8394–8405.
149. Wang, H.; Tang, H.; Shi, S.; Li, A.; Li, Z.; Schiele, B.; Wang, L. Unitr: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Melbourne, Australia, 2–6 October 2023; pp. 6792–6802.
150. Cha, J.; Joo, M.; Park, J.; Lee, S.; Kim, I.; Kim, H.J. Robust Multimodal 3D Object Detection via Modality-Agnostic Decoding and Proximity-Based Modality Ensemble. *arXiv* **2024**, arXiv:2407.19156.
151. Kim, Y.; Park, K.; Kim, M.; Kum, D.; Choi, J.W. 3D Dual-Fusion: Dual-Domain Dual-Query Camera-LIDAR Fusion for 3D Object Detection. *arXiv* **2022**, arXiv:2211.13529.
152. Koh, J.; Lee, J.; Lee, Y.; Kim, J.; Choi, J.W. Mgtanet: Encoding Sequential Lidar Points Using Long Short-Term Motion-Guided Temporal Attention for 3d Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1179–1187.
153. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3d Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, online, 19–25 June 2021; pp. 11784–11793.
154. Zhu, X.; Ma, Y.; Wang, T.; Xu, Y.; Shi, J.; Lin, D. Ssn: Shape Signature Networks for Multi-Class Object Detection from Point Clouds. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXV 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 581–597.
155. Shao, H.; Wang, L.; Chen, R.; Waslander, S.L.; Li, H.; Liu, Y. Reasonnet: End-to-End Driving with Temporal and Global Reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13723–13733.
156. Shao, H.; Wang, L.; Chen, R.; Li, H.; Liu, Y. Safety-Enhanced Autonomous Driving Using Interpretable Sensor Fusion Transformer. In Proceedings of the Conference on Robot Learning, PMLR, Tokyo, Japan, 11–15 December 2023; pp. 726–737.
157. Wu, P.; Jia, X.; Chen, L.; Yan, J.; Li, H.; Qiao, Y. Trajectory-Guided Control Prediction for End-to-End Autonomous Driving: A Simple yet Strong Baseline. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6119–6132.
158. Chen, D.; Krähenbühl, P. Learning from All Vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 17222–17231.
159. Chitta, K.; Prakash, A.; Jaeger, B.; Yu, Z.; Renz, K.; Geiger, A. Transfuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 12878–12895. [CrossRef]
160. Renz, K.; Chen, L.; Marcu, A.-M.; Hünermann, J.; Hanotte, B.; Karnsund, A.; Shotton, J.; Arani, E.; Sinavski, O. CarLLaVA: Vision Language Models for Camera-Only Closed-Loop Driving. *arXiv* **2024**, arXiv:2406.10165.
161. Jaeger, B.; Chitta, K.; Geiger, A. Hidden Biases of End-to-End Driving Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Melbourne, Australia, 2–6 October 2023; pp. 8240–8249.
162. Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; Carion, N. Mdetr-Modulated Detection for End-to-End Multi-Modal Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, online, 11–17 October 2021; pp. 1780–1790.

163. Deruyttere, T.; Grujicic, D.; Blaschko, M.B.; Moens, M.-F. Talk2Car: Predicting Physical Trajectories for Natural Language Commands. *IEEE Access* **2022**, *10*, 123809–123834. [CrossRef]
164. Deruyttere, T.; Vandenhende, S.; Grujicic, D.; Liu, Y.; Van Gool, L.; Blaschko, M.; Tuytelaars, T.; Moens, M.-F. Commands 4 Autonomous Vehicles (C4av) Workshop Summary. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–26.

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