Review

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

Michał Kozłowski *, Szymon Racewicz, Sławomir Wierzbicki *

Posted Date: 24 July 2024

doi: 10.20944/preprints202407.1857.v1

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

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 the latest state-of-the-art AI solutions employed in image analysis for AVs. Datasets such as Cityscapes, NuScenes, and CARLA 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), Recurrent Neural Networks (RNNs), Transformer models, and Generative Adversarial Networks (GANs), 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, and vehicle control in virtual environments. Simultaneously, the role of multisensor datasets and simulation platforms like AirSim, TORCS, and SUMMIT in enriching the training data and testing environments for AVs is highlighted. By synthesizing information on datasets, AI solutions, and comparative performance evaluations, the 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

1. Introduction

Environmental image analysis is pivotal in the realm of 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. Autonomous vehicles, i.e. cars [1–3], ships [4,5], underwater vehicles [6–8], unmanned aerial vehicles (UAVs) [9–11] or 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.

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) [18]. 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. In Levels 1 (Driver Assistance) and 2 (Partial Automation), image analysis becomes more critical, managing tasks such as adaptive cruise control and 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. 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

 doi:10.20944/preprints202407.1857.v1

vehicle handling all driving aspects in any condition without human oversight, requiring image analysis capabilities surpassing human abilities for interpreting visual information and making instantaneous decisions.

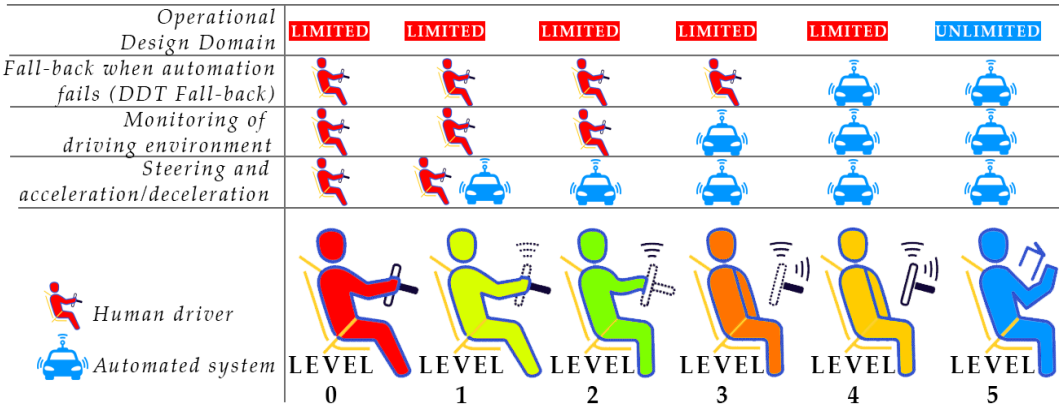


Figure 1. SAE J3016 Levels of Driving Automation.

The significance of image analysis spans all autonomy levels, enhancing situational awareness, safety and redundancy, complex decision-making, adaptability, precise mapping and localization, and ultimately supporting the regulatory and social acceptance of autonomous systems.

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 [19,20]. Real-time image processing methods are designed for road detection, obstacle recognition, traffic light detection or even number plate recognition [21–23]. Road detection techniques utilize similarity analysis and structural similarity index analysis to identify road surfaces in grid image areas [24]. Object detection and tracking remain challenging tasks, with image classification algorithms such as CNNs playing a pivotal role in steering the vehicle [25–27].

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.

Additionally, image analysis is crucial for the vehicle's adaptability to changing environmental conditions. It helps in identifying and differentiating between various surfaces and obstacles, even under adverse conditions such as poor lighting, inclement weather, or complex urban landscapes. Such adaptability is critical for the reliable operation of autonomous vehicles across diverse real-world environments.

Moreover, environmental image analysis improves the efficiency of autonomous vehicles by enabling accurate environmental mapping and behavior prediction of other road users. This capability optimizes routes, reduces travel time, and enhances fuel efficiency.

Continuous advancements in image processing technologies and artificial intelligence are propelling the evolution of autonomous vehicles. Improvements in deep learning algorithms and computational power yield more precise and rapid image analysis, enhancing the performance and reliability of autonomous driving systems.

In summary, environmental image analysis is integral to the functionality and progression of autonomous vehicles. It underpins the vehicle's ability to accurately perceive its surroundings, ensuring safety, enhancing situational awareness, providing adaptability to diverse conditions, and improving operational efficiency. As technology evolves, the role of image analysis in autonomous vehicles will continue to grow, fostering safer and more efficient transportation solutions.

 doi:10.20944/preprints202407.1857.v1

2. Aims of the Article

The primary aim of this article is to provide a comprehensive overview of the available datasets and the latest artificial intelligence (AI) solutions utilized for image analysis in the context of autonomous vehicles. This encompasses a detailed survey of the most significant and widely-used datasets that serve as benchmarks for training and evaluating AI models, such as Cityscapes [28], NuScenes [29], and CARLA [30]. Each dataset's unique characteristics, including the diversity of scenarios, sensor configurations, and annotation quality, will be explored to underscore their relevance and applicability to different aspects of autonomous driving.

Simultaneously, the article aims to illuminate the cutting-edge AI technologies that drive advancements in image analysis. It will delve into the intricacies of state-of-the-art deep learning architectures, including but not limited to Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformer models, and Generative Adversarial Networks (GANs). There will also be new approaches that use less intuitive approaches, such as deep neural models inspired by the PID controller [31]. The exploration will cover recent breakthroughs in these domains, such as innovations in transfer learning, multitask learning, and unsupervised learning methodologies. By analyzing the latest research findings and technological developments, the article will highlight how these AI solutions enhance the environmental perception capabilities of autonomous vehicles, facilitating tasks like semantic segmentation, 3d object detection and trajectory prediction.

Ultimately, by synthesizing information on both datasets and AI solutions, the article aims to provide a valuable resource for researchers, developers, and industry stakeholders, offering insights into the current landscape and future directions of image analysis in autonomous vehicle technology.

The second purpose of this article is to conduct a comparative analysis of the effectiveness and application of various AI techniques in real-world scenarios, with a particular focus on three critical tasks: semantic image segmentation, 3D object detection in video, and vehicle control in virtual environments. This involves detailed examination of how different AI methodologies perform across these key functions, which are essential for the robust operation of autonomous vehicles.

For semantic image segmentation, several deep learning models have been compared, including VLTSeg [32], PIDNet [31], Semantic Flow [33], and Rethinking Dilated Convolution [34]. The comparison will consider metrics such as the average intersection across the Union (mIoU) and the models' performance in terms of processing speed, measured by frames per second (fps). These evaluations were conducted to find a balance between accuracy and performance.

In the domain of 3D object detection in video, the article will compare approaches such as EA-LSS, CenterPoint, and other state-of-the-art networks. These techniques will be assessed on their ability to accurately detect and localize objects in three-dimensional space. Factors such as detection range, precision, robustness, and scalability will be examined to determine the best-suited techniques for practical deployment.

For vehicle control in virtual environments, attention will be given to reinforcement learning algorithms and simulation-based decision-making frameworks. The article will analyze how algorithms like ReasonNet [35], InterFuser [36], TCP [37] and others that have achieved leading results in selected tests. Success metrics will include the systems' learning efficiency, driving score, route completion and infraction penalty.

By integrating empirical data from simulations and field trials, the article will provide a comprehensive comparison of these AI techniques, identifying their strengths, weaknesses, and suitability for various applications in autonomous driving. This insight aims to inform ongoing research and development, guiding the implementation of the most effective AI strategies to enhance the performance and reliability of autonomous vehicles in real-world scenarios.

3. AI in Autonomous Vehicles

Research on image analysis for autonomous vehicles is a vibrant and rapidly evolving field that encompasses numerous interdisciplinary areas, including computer vision, machine learning, sensor fusion, and robotics. Existing literature and publications cover a wide range of topics, from algorithm development to practical applications and system validation. Object detection and classification are

 doi:10.20944/preprints202407.1857.v1

fundamental tasks for autonomous vehicles, enabling the identification of pedestrians, vehicles, traffic signs, and other road users [38]. Notable research includes YOLO (You Only Look Once), proposed by Redmon et al., known for its efficiency and accuracy in real-time object detection [39–41]. Additionally, R-CNN (Region-Based Convolutional Neural Networks) [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.

Depth estimation and stereo vision are crucial for understanding spatial relationships. Notable advances include traditional stereo matching techniques like Semi-Global Matching (SGM) [45] by Hirschmüller, used in stereo vision systems for calculating depth maps, and monocular depth estimation using deep learning by Eigen et al. [46], which significantly advance the capability to estimate depth from single images. Sensor fusion combines data from multiple sensor types such as cameras, LiDAR, and radar to enhance perception system robustness. Advances in deep learning for sensor fusion are evaluated systematically in projects like KITTI [20,47], which combine 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" [48] discusses neural networks' vulnerability to small, intentionally crafted perturbations and proposes strategies for improving robustness. Safety and verification platforms like the CARLA [30], TORCS [49] simulators provide rigorous testing and validation environments.

Continuous development of better image analysis algorithms and architectures is paramount. Notable advancements include EfficientNet by the Google Brain team [50], which uses a compound scaling method for state-of-the-art results in efficiency and accuracy, and Vision Transformers (ViTs) that demonstrate promise in scaling and capturing long-range dependencies. Generative Adversarial Networks (GANs) have emerged as a powerful tool in image analysis, significantly impacting the domain of autonomous vehicles [51]. Introduced by Ian Goodfellow and colleagues in 2014, GANs consist of a generator network that creates realistic data samples and a discriminator network that distinguishes between real and generated data. This adversarial training process allows GANs to generate highly realistic images, which have various applications in image analysis for autonomous vehicles [52–54].

GANs are used to create diverse and realistic images, augmenting training datasets for object detection, semantic segmentation, and other tasks, thereby improving model robustness and generalizability [55]. They also simulate a wide range of driving conditions, generating images illustrating various weather scenarios, lighting conditions, and road environments, which allow autonomous vehicle systems to train on a broader spectrum of scenarios without extensive real-world data collection. GANs like SRGAN (Super-Resolution GAN) enhance the resolution of images captured by vehicle cameras, improving clarity and detail vital for accurate object detection and scene understanding [56]. Additionally, GANs can reduce noise in images captured under low-light or noisy conditions, further improving image quality. Techniques such as CycleGAN enable translating images from one domain to another (e.g., from synthetic to real-world images), crucial for training models in simulated environments and applying them in real-world scenarios [57]. Style transfer via GANs allows models trained on one type of imagery (e.g., daytime) to perform well in another type (e.g., nighttime), by transferring styles between domains.

GANs enhance the accuracy of semantic segmentation tasks. Models like SegAN combine segmentation networks with adversarial training to achieve fine-grained results. Conditional GANs (cGANs) generate images conditioned on specific labels or inputs, such as generating segmented versions of real-world driving images, which is useful for training and validating segmentation

 doi:10.20944/preprints202407.1857.v1

models [58]. GANs create highly realistic virtual driving scenarios for testing and validating autonomous driving algorithms. They simulate rare and dangerous driving scenarios safely, such as sudden pedestrian crossings or abrupt weather changes, enabling extensive testing without real-world risks. GANs are employed to generate adversarial examples to test and improve the robustness of perception systems against intentionally perturbed inputs, leading to the development of more robust neural networks. GANs predict future frames in video sequences, aiding in understanding the motion of objects and predicting their future positions—crucial for autonomous navigation [59]. They can also synthesize motion from one scene into another, enhancing the vehicle's ability to understand dynamic environments.

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

4. 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 a realistic simulation. 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 [60] 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.

4.1. Image Collections

The Table 1 enumerates various image datasets crucial for training deep neural models in autonomous driving. The Cityscapes dataset emerges as the most frequently cited (3,411 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.

 doi:10.20944/preprints202407.1857.v1

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

Name	Description	Cited	Ref.
Cityscapes	Provides semantic, instance-wise, and dense pixel annotations for 30 classes grouped into 8 categories.	3,411	[28]
Waymo	High resolution sensor data collected by Waymo Driver in various conditions.	398	[61]
KITTI-360	Popular KITTI dataset with comprehensive semantic/instance labels in 2D and 3D.	181	[47]
IDD	Road scene understanding in unstructured environments dataset.	90	[62]
INTERACTION	Contains naturalistic motions of traffic participants in highly interactive scenarios.	73	[63]
SemanticPOSS	3D semantic segmentation dataset collected in Peking University.	60	[64]
WoodScape	Extensive fisheye camera automotive dataset with nine tasks and 40 classes annotations.	49	[65]
Lost and Found	Lost-cargo image sequence dataset with pixelwise annotations of obstacles and free-space.	47	[66]
DrivingStereo	Over 180k images for stereo vision, larger than KITTI Stereo dataset.	42	[67]
Fishyscapes	Evaluates pixel-wise uncertainty estimates towards detecting anomalous objects.	44	[68]
ROAD Anomaly	Contains images of unusual dangers encountered by vehicles, such as animals and traffic cones.	44	[69]
PandaSet	Dataset captured with high-precision autonomous vehicle sensor kit.	39	[70]
KITTI Road	Road and lane estimation benchmark.	38	[71]
Talk2Car	Cross-disciplinary dataset for grounding natural language into visual space.	38	[72]
MVSEC	Data collection designed for developing 3D perception algorithms for event-based cameras.	26	[73]
KAIST Urban	Raw sensor data for vehicle navigation with development tools in the ROS environment.	19	[74]
Cityscapes 3D	Extends the original Cityscapes dataset with 3D bounding box annotations for vehicles.	10	[28]
RailSem19	Dataset for semantic rail scene understanding with images from rail vehicles.	8	[75]
RoadAnomaly21	Contains images with at least one anomalous object such as animals or unknown vehicles.	8	[76]
EuroCity Persons	Annotations of pedestrians, cyclists, and riders in urban traffic scenes from 31 cities in Europe.	6	[77]
DOLPHINS	Dataset for testing vehicle-to-everything (V2X) network in autonomous driving.	5	[78]
PSI	Dataset capturing dynamic intent changes for pedestrians crossing in front of ego-vehicles.	5	[79,80]
TICaM	Dataset for vehicle interior monitoring using a wide-angle depth camera.	5	[81]
Zenseact	Dataset collected over 2 years across 14 European countries with full sensor suite.	5	[82]
OoDIS	Dataset for anomaly instance segmentation in autonomous driving.	4	[83]
LOOK	Real-world scenarios for autonomous vehicles focusing on pedestrian interactions.	3	[84]

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.

 doi:10.20944/preprints202407.1857.v1

Anomaly instance segmentation is addressed by OoDIS, while real-world interaction scenarios are captured in LOOK (3 citations). The diversity and thematic grouping of these datasets underscore their relevance and application in developing robust autonomous vehicle systems.

4.2. Video Stocks

Table 2 lists various video datasets used for training deep neural models in autonomous driving, highlighting key datasets and their applications. The nuScenes dataset stands out as the most widely cited (1,695 citations) with its comprehensive suite of 32-beam LiDAR, six cameras, and radars providing full 360° coverage, crucial for holistic vehicle perception.

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

Name	Description	Cited	Ref.
nuScenes	Full autonomous vehicle data suite: 32-beam LiDAR, 6 cameras and radars with complete 360° coverage.	1,695	[29]
Virtual KITTI	Photo-realistic synthetic video dataset for several video understanding tasks.	124	[85]
CULane	Lane detection dataset collected by cameras mounted on six different vehicles in Beijing.	77	[86]
ApolloScape	Large dataset with over 140,000 video frames from various locations in China.	68	[87]
ROAD	Tests an autonomous vehicle's ability to detect road events using annotated videos.	21	[88]
V2V4Real	Data collected by two vehicles equipped with multi-modal sensors driving together through diverse scenarios.	17	[89]
TITAN	700 labeled video-clips with odometry captured from a moving vehicle in Tokyo.	12	[90]
BrnoCompSpeed	Vehicles annotated with precise speed measurements from LiDAR and GPS tracks.	11	[91]
CCD	Real traffic accident videos captured by dashcam with diverse annotations.	10	[92]
BLVD	Large scale 5D semantics dataset collected in China's Intelligent Vehicle Proving Center.	9	[93]
HEV-I	Dataset includes video clips of real human driving in different intersections in the San Francisco Bay Area.	5	[94]
Ford CVaL	Dataset collected by an autonomous ground vehicle testbed equipped with multiple sensors, collected in Michigan.	3	[95]
TbV Dataset	Over 1000 scenarios captured by autonomous vehicles, each log represents a continuous observation of a scene around a self-driving vehicle.	2	[96]
Argoverse 2 Map Change	Temporal annotations indicating map changes within 30 meters of an autonomous vehicle.	1	[60]
D2CITY	Large-scale collection of dashcam videos collected by vehicles on DiDi's platform.	1	[97]
DADE	Sequences acquired by agents (ego vehicles) within a 5-hour time frame, totaling 990k frames.	1	[98]
DMPD	Test set contains images and pedestrian labels captured from a vehicle during a 27-minute drive.	1	[99]
INDRA	Dataset consisting of 104 videos with annotated road crossing safety labels and vehicle bounding boxes.	1	[100]
METEOR	Consists of 1000+ one-minute video clips with annotated frames and bounding boxes for surrounding vehicles and traffic agents.	1	[101]
METU-VIREF	VIRAT dataset for surveillance containing primarily people and vehicles, aligned with videos from the ILSVRC dataset.	1	[102]
RoadTextVQA	Video question answering dataset for in-vehicle conversations.	1	[103]
TLV	Real-world datasets based on NuScenes and Waymo for temporal logic.	1	[104]
Vehicle-Rear	Dataset for vehicle identification with high-resolution videos, including make, model, color, and license plates.	1	[105]
IMO	Contains images, stereo disparity, and vehicle labels with ground truth annotations.	0	[106]
LISA Vehicle Detection	Dataset for vehicle detection with video sequences captured at different times and varying traffic conditions.	0	[107]

 doi:10.20944/preprints202407.1857.v1

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 TbV Dataset with over 1,000 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.

4.3. Simulators

Table 3 presents various simulators designed to train deep neural models for autonomous driving. The most widely cited simulator is CARLA (1,128 citations), offering a versatile urban driving simulation with detailed annotations including 12 semantic classes, bounding boxes, and vehicle measurements.

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.	1,128	[30]
AirSim	Simulator for drones, cars, and more, built on Unreal Engine, with support for SIL and HIL.	248	[108]
TORCS	Driving simulator capable of simulating elements of vehicular dynamics.	91	[49]
V2X-SIM	Synthetic collaborative perception dataset for autonomous driving, collected from both roadside and vehicles.	17	[109]
SUMMIT	Supports a wide range of applications including perception, control, planning, and end-to-end learning.	13	[110]
CVRPTW	Instances of the Capacitated Vehicle Routing Problem with Time Windows for various customer nodes.	7	[111]
CARL	Control suite extended with physics context features for AI training.	6	[112]
3D VTSim	Dataset collected using driving simulation for accurate 3D bounding box annotations.	4	[113]
MUAD	Dataset with realistic synthetic images under diverse weather conditions, annotated for multiple tasks.	3	[114]
SDN	Navigation benchmark with trials and control streams developed to evaluate dialogue moves and physical navigation actions.	2	[115]
MULTIROTOR-GYM	Multirotor gym environment for learning control policies for UAVs.	2	[116,117]

 doi:10.20944/preprints202407.1857.v1

DEAP CITY	City pollution data including daily pollutant and meteorological features, alongside total vehicle mileage.	1	[118]
EviLOG	Real-world lidar point clouds from a test vehicle with the same lidar setup as simulated lidar.	1	[119]

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 [120].

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 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 simulation.

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

5. Comparison of Selected AI Models

Comparative testing of various models on selected datasets is critical for evaluating and understanding the performance 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.

5.1. 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.

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

Figure 2. An example of semantic segmentation on the Cityscapes set where: (a) Original image; (b) ground truth image; (c) model prediction result [121].

 doi:10.20944/preprints202407.1857.v1

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.
VLTSeg	86.4%	76(n/a)	2023	[32]
PIDNet-L	80.6%	31.1 (3090)	2022	[31]
SFNet-R18	80.4%	25.7 (1080Ti)	2020	[33]
PIDNet-M	79.8%	42.2 (3090)	2022	[31]
PIDNet-S	78.6%	93.2 (3090)	2022	[31]
RegSeg	78.3%	30 (n/a)	2021	[34]
PP-LiteSeg-B2	77.5%	102.6 (1080Ti)	2022	[122]
DDRNet-23-slim	77.4%	101.6 (2080Ti)	2021	[123]
STDC2-75	76.8%	97.0 (1080Ti)	2021	[124]
U-HarDNet-70	75.9%	53 (1080Ti)	2019	[125]
HyperSeg-M	75.8%	36.9(n/a)	2020	[126]
SwiftNetRN-18	75.5%	39.9(n/a)	2019	[127]
STDC1-75	75.3%	126.7(n/a)	2021	[124]
BiSeNet V2-Large	75.3%	47.3(n/a)	2021	[128]
TD4-BISE18	74.9%	47.6 (Titan X)	2020	[129]
PP-LiteSeg-T2	74.9%	143.6 (1080Ti)	2022	[122]

The VLTSeg model, achieving the highest mIoU of 86.4%, leverages visual language representations from the CLIP framework to bridge domain gaps in semantic segmentation. By integrating pre-trained multimodal features, VLTSeg enhances segmentation performance across diverse environments. However, its fps data on GPU remains unspecified.

PIDNet variants (PIDNet-L, PIDNet-M, and PIDNet-S) demonstrate strong performance with mIoUs surpassing 78%, balancing accuracy and speed. Notably, PIDNet-S achieves 93.2 fps on an Nvidia RTX 3090. PIDNet's architecture, inspired by Proportional-Integral-Derivative (PID) controllers, integrates distinct branches to enhance feature extraction and aggregation, optimizing both spatial detail preservation and contextual information assimilation for real-time applications.

SFNet-R18 achieves an mIoU of 80.4% with 25.7 fps on a 1080Ti GPU, utilizing a flow-guided feature aggregation mechanism that combines local and global contexts to improve segment coherence and inference speed. RegSeg mirrors SFNet-R18's performance (78.3% mIoU, 25.7 fps on 1080Ti) by optimizing dilation rates within convolutional layers, balancing receptive field enlargement and computational efficiency to preserve fine-grained details while maintaining real-time processing speeds.

The PP-LiteSeg model distinguishes itself with high operation speed, achieving up to 143.6 fps on a 1080Ti GPU, while maintaining a decent mIoU range (74.9%-77.5%). It utilizes an efficient backbone and a lightweight decoder to enhance both segmentation accuracy and processing speed. Similarly, DDRNet-23-slim (77.4% mIoU, 101.6 fps on 2080Ti) employs a dual-resolution strategy to optimize spatial precision and contextual understanding.

STDC2-75 (76.8% mIoU, 97.0 fps on 1080Ti) and STDC1-75 (75.3% mIoU, 126.7 fps) models integrate a dual-branch design, balancing spatial detail retention and contextual information assimilation. HarDNet (75.9% mIoU, 53 fps on 1080Ti) utilizes a decomposed network structure and strategically spaced shortcut connections to reduce memory access costs while maintaining computational efficiency. HyperSeg (75.8% mIoU, 36.9 fps) employs a patch-wise hypernetwork approach for dynamically generating lightweight, patch-specific segmentation models.

 doi:10.20944/preprints202407.1857.v1

BiSeNet V2-Large (75.3% mIoU, 47.3 fps) introduces a bilateral network architecture with guided aggregation to balance spatial precision and contextual understanding, while TD4-BISE18 (74.9% mIoU, 47.6 fps on Titan X) captures temporal dependencies through a mechanism that distributes computation across frames, reducing redundant processing.

In conclusion, while models like VLTSeg and PIDNet variants excel in segmentation accuracy, others such as PP-LiteSeg and STDC prioritize real-time performance, making them suitable for applications requiring rapid processing speeds. Each model employs unique innovations to achieve high performance with minimal hardware resources, demonstrating the diverse approaches to efficient semantic segmentation.

5.2. 3D Objects Detection

Detection of 3D objects on 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 [130].

The performance of various 3D object detection models evaluated on the nuScenes dataset is presented in Table 5. The models are compared based on several metrics: 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), mAAE (mean Average Attribute Error), along with the year of publication and reference.

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	[131]
BEVFusion-e	0.76	0.75	0.24	0.23	0.32	0.22	0.13	2022	[132]
FocalFormer3D-F	0.75	0.72	0.25	0.24	0.33	0.23	0.13	2023	[133]
UniTR	0.75	0.71	0.24	0.23	0.26	0.24	0.13	2023	[134]
FocalFormer3D-TTA	0.74	0.71	0.24	0.24	0.32	0.20	0.13	2023	[133]
3D Dual-Fusion_T	0.73	0.71	0.26	0.24	0.33	0.27	0.13	2022	[135]
FocalFormer3D-L	0.73	0.69	0.25	0.24	0.34	0.22	0.13	2023	[133]
MGTANet	0.73	0.67	0.25	0.23	0.31	0.19	0.12	2022	[136]
CenterPoint	0.71	0.67	0.25	0.24	0.35	0.25	0.14	2020	[137]
SSN	0.62	0.51	0.34	0.24	0.43	0.27	0.09	2020	[138]

The EA-LSS model demonstrates the highest performance with NDS 0.78 and mAP 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.

 doi:10.20944/preprints202407.1857.v1

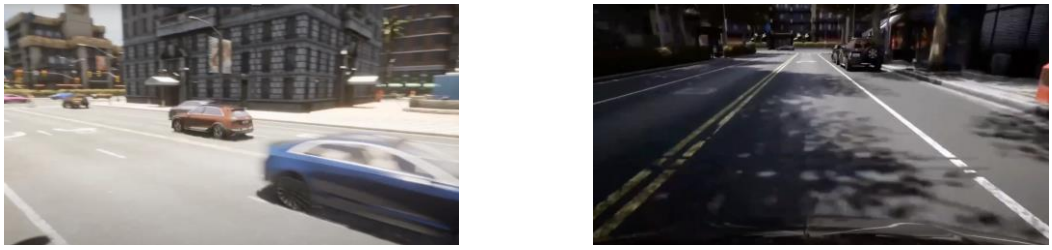
Other notable models include BEVFusion, which integrates multi-sensor data into a unified BEV representation to support concurrent tasks such as detection, segmentation, and tracking. FocalFormer3D employs a targeted mechanism to prioritize challenging instances within the 3D space, leveraging focal attention combined with transformer networks for 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.

5.3. 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). She evaluated algorithms 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 8 RGB cameras, 2 LIDAR scanners and 4 radars.



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

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

Tables 6 and 7 present the performance of 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	[35]
InterFuser	76.18	88.23	0.84	2022	[36]
TGCP	75.14	85.63	0.87	2022	[37]
LAV	61.84	94.46	0.64	2022	[139]
TransFuser	61.18	86.70	0.71	2022	[140]
TransFuser(Reproduced)	55.04	89.65	0.63	2022	[140]
TGCP(Reproduced)	47.91	65.73	0.77	2023	[37]
Latent TransFuser	45.20	66.31	0.72	2022	[140]

 doi:10.20944/preprints202407.1857.v1

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

LAV demonstrates the highest Route Completion score (94.46) and the 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 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 biases in autonomous driving systems, revealing how 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.

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	[141]
CarLLaVA(Map Track)	6.25	18.89	0.39	2024	[141]
TF++(Map Track)	5.56	11.82	0.47	2024	[142]
TF++	5.18	11.34	0.48	2024	[142]

In **Table 7** the 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 the lowest Infraction Score (0.39), making it exceptionally robust in route execution and safety.

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.

The comparative analysis reveals that models like ReasonNet and LAV from CARLA Leaderboard 1.0 demonstrate 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 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.

6. Discussion, Limitations and Future Research Trends

6.1. Discussion

The convergence of image analysis technology and autonomous vehicles (AVs) is catalyzing profound advancements in self-driving capabilities. This paper has elucidated a spectrum of innovative AI methodologies, namely Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and transformer models, illuminating their extensive use in the perception systems of AVs. These models have showcased remarkable proficiency in tasks such as semantic segmentation, 3D object detection, and trajectory prediction, thereby enhancing the decision-making abilities of AVs.

However, while CNNs and their variants remain pivotal in image analysis due to their robustness in feature extraction, transformer models present a significant leap forward by offering a

 doi:10.20944/preprints202407.1857.v1

more integrated approach to multi-sensor data fusion. GANs and other novel architectures like the dual-query fusion approach further advance capabilities in object detection by leveraging the complementary strengths of LiDAR and camera data. The empirical results from benchmarks, especially datasets like Cityscapes, nuScenes, and CARLA, validate these technologies' efficacy in diverse and challenging driving scenarios.

In comparing various models, VLTSeg and PIDNet were shown to be exemplary in maintaining a balance between accuracy and performance. In 3D detection, EA-LSS, BEVFusion, and CenterPoint exhibited significant promise by integrating multi-modal data and enhancing object boundary delineation and spatial accuracy. For vehicle control in virtual environments, ReasonNet and CarLLaVA demonstrated high route completion efficiency and low infraction rates, attributed to their integration of visual and language data for decision-making processes.

6.2. Limitations

Despite these advances, several limitations remain impeding the full realization of autonomous driving:

1. Annotated datasets

While expansive datasets like Cityscapes and nuScenes render robust model training, the necessity for substantial human annotation remains a bottleneck. Data labels must be precise, consistent, and extensive to ensure comprehensive model training, yet the manual effort required is both time-consuming and expensive.

2. Generalization

Models trained on specific datasets often struggle to generalize effectively across different environments and conditions. Differences in weather, road conditions, and regions can impact the performance of perception systems, necessitating continual data collection and model retraining.

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

The 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.

6.3. 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.

 doi:10.20944/preprints202407.1857.v1

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 novel sensor data

Expanding the use of multimodal sensors, including radar and advanced LiDAR systems, coupled with innovative data fusion algorithms, can enhance the perception capabilities. Collaborative perception and V2X (Vehicle-to-Everything) communication technologies represent promising avenues for research.

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.

7. Conclusions

The integration of advanced AI methodologies within the perception systems of AVs has driven significant progress in the realm of self-driving technology. This paper has extensively reviewed foundational and state-of-the-art models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and transformer models. These models exhibit remarkable capabilities across various critical tasks, including semantic segmentation, 3D object detection, and trajectory prediction, thereby substantially elevating the decision-making efficacy of AVs.

Among these methodologies, CNNs continue to play a crucial role due to their robustness in feature extraction. However, transformer models have emerged as a pivotal advancement, offering superior data fusion capabilities and showing promise in integrating multi-sensor data. GANs, alongside novel dual-query fusion approaches, are pushing boundaries further by effectively combining LiDAR and camera data to enhance object detection accuracy and spatial precision. Empirical validations using benchmarks such as Cityscapes, nuScenes, and CARLA underline the effectiveness of these methodologies across a spectrum of driving scenarios and conditions. Models such as VLTSeg and PIDNet stand out for their balance of accuracy and performance in image segmentation, while EA-LSS, BEVFusion, and CenterPoint demonstrate significant potential in 3D detection tasks. Furthermore, ReasonNet and CarLLaVA have shown high efficacy in vehicle control within virtual environments by integrating visual and language data, leading to improved route completion and lower infraction rates.

In conclusion, while remarkable strides have been made in the application of AI for image analysis in autonomous vehicles, ongoing research and development are imperative for overcoming current barriers. Such continuous innovation will pave the way for more advanced, reliable, and safe autonomous driving systems, ultimately bringing us closer to a future where fully autonomous vehicles are an integral part of everyday life.

Author Contributions: Conceptualization, M. Kozłowski; methodology, M. Kozłowski and S. Racewicz; formal analysis, M. Kozłowski and S. Racewicz; investigation, M. Kozłowski and S. Racewicz; resources, M. Kozłowski and S. Racewicz; writing—original draft preparation, M. Kozłowski and S. Racewicz; writing—review and editing, M. Kozłowski and S. Racewicz; visualization, M. Kozłowski and S. Racewicz; supervision, S. Wierzbicki; project administration, S. Wierzbicki. 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. *Applied Sciences* **2024**, *14*, 5453, doi:10.3390/app14135453.
2. Schrader, M.; Hainen, A.; Bittle, J. Extracting Vehicle Trajectories from Partially Overlapping Roadside Radar. *Sensors* **2024**, *24*, 4640, doi:10.3390/s24144640.
3. Booth, L.; Karl, C.; Farrar, V.; Pettigrew, S. Assessing the Impacts of Autonomous Vehicles on Urban Sprawl. *Sustainability* **2024**, *16*, 5551, doi:10.3390/su16135551.
4. Muhovič, J.; Perš, J. Correcting Decalibration of Stereo Cameras in Self-Driving Vehicles. *Sensors (Switzerland)* **2020**, *20*, 1–17, doi:10.3390/s20113241.
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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *PP*, 1–14, doi:10.1109/jstars.2024.3402540.
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. *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering - OMAE* **2019**, *7B-2019*, doi:10.1115/OMAE2019-95321.
7. Nian, R.; Liu, F.; He, B. An Early Underwater Artificial Vision Model in Ocean Investigations via Independent Component Analysis. *Sensors (Basel, Switzerland)* **2013**, *13*, 9104–9131, doi:10.3390/s130709104.
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, doi:10.3390/s111110958.
9. Kim, J.; Cho, J. Rgdinet: Efficient Onboard Object Detection with Faster r-Cnn for Air-to-Ground Surveillance. *Sensors* **2021**, *21*, 1–16, doi:10.3390/s21051677.
10. Salles, R.N.; Velho, H.F. de C.; Shiguemori, E.H. Automatic Position Estimation Based on Lidar × Lidar Data for Autonomous Aerial Navigation in the Amazon Forest Region. *Remote Sensing* **2022**, *14*, 1–27, doi:10.3390/rs14020361.
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 Sensing* **2018**, *10*, 1–31, doi:10.3390/rs10111829.
12. Wang, H.; Lu, E.; Zhao, X.; Xue, J. Vibration and Image Texture Data Fusion-Based Terrain Classification Using WKNN for Tracked Robots. *World Electric Vehicle Journal* **2023**, *14*, 1–14, doi:10.3390/wevj14080214.
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*, 1–20, doi:10.3390/math9233139.
14. Ci, W.; Huang, Y. A Robust Method for Ego-Motion Estimation in Urban Environment Using Stereo Camera. *Sensors (Switzerland)* **2016**, *16*, 1–14, doi:10.3390/s16101704.
15. Kim, B.J.; Lee, S.B. A Study on the Evaluation Method of Autonomous Emergency Vehicle Braking for Pedestrians Test Using Monocular Cameras. *Applied Sciences (Switzerland)* **2020**, *10*, 1–15, doi:10.3390/app10134683.
16. Kim, Y.-W.; Byun, Y.-C.; Krishna, A.V. Portrait Segmentation Using Ensemble of Heterogeneous Deep-Learning Models. *Entropy* **2021**, *23*, 197.
17. Kim, J. Detection of Road Images Containing a Counterlight Using Multilevel Analysis. *Symmetry* **2021**, *13*, doi:10.3390/sym13112210.
18. International, S. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. *SAE international* **2018**, *4970*, 1–5.

19. Wang, Y.F. Computer Vision Analysis for Vehicular Safety Applications. In Proceedings of the Proceedings of the International Telemetering Conference; International Foundation for Telemetering, January 1 2015; Vol. 82, pp. 944–953.
20. Yebes, J.J.; Bergasa, L.M.; García-Garrido, M.Á. Visual Object Recognition with 3D-Aware Features in KITTI Urban Scenes. *Sensors (Switzerland)* **2015**, *15*, 9228–9250, doi:10.3390/s150409228.
21. 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, December 6 2021; pp. 1–4.
22. Kumawat, K.; Jain, A.; Tiwari, N. Relevance of Automatic Number Plate Recognition Systems in Vehicle Theft Detection †. *Engineering Proceedings* **2023**, *59*, doi:10.3390/engproc2023059185.
23. 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*, doi:10.3390/math12081206.
24. 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, November 6 2020; pp. 63–66.
25. A, S.I.; R, K.; Shanmugasundaram, H.; A, B. prasad; R, K.; J, M.B. Lane Detection Using Deep Learning Approach. In Proceedings of the 2022 1st International Conference on Computational Science and Technology (ICCST); IEEE, November 9 2022; pp. 945–949.
26. 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 (Switzerland)* **2021**, *10*, 1–21, doi:10.3390/electronics10111266.
27. Itu, R.; Danescu, R. Fully Convolutional Neural Network for Vehicle Speed and Emergency-Brake Prediction. *Sensors* **2024**, *24*, doi:10.3390/s24010212.
28. 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 Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; pp. 3213–3223.
29. 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; 2020.
30. Nikolenko, S.I. Synthetic Data for Deep Learning. *CoRR* **2019**, *abs/1909.11512*.
31. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023; pp. 19529–19539.
32. 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 preprint arXiv:2312.02021* **2023**.
33. 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, August 23–28, 2020, Proceedings, Part I 16; Springer, 2020; pp. 775–793.
34. Gao, R. Rethinking Dilated Convolution for Real-Time Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; pp. 4675–4684.
35. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023; pp. 13723–13733.
36. 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, 2023; pp. 726–737.
37. 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. *Advances in Neural Information Processing Systems* **2022**, *35*, 6119–6132.
38. 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.
39. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2016.

40. Yao, C.; Liu, X.; Wang, J.; Cheng, Y. Optimized Design of EdgeBoard Intelligent Vehicle Based on PP-YOLOE+. *Sensors* **2024**, *24*, 3180, doi:10.3390/s24103180.
41. 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.
42. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-Cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision; 2017; pp. 2961–2969.
43. Feldsar, B.; Mayer, R.; Rauber, A. Detecting Adversarial Examples Using Surrogate Models. *Machine Learning and Knowledge Extraction* **2023**, *5*, 1796–1825.
44. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *39*, 2481–2495.
45. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on pattern analysis and machine intelligence* **2007**, *30*, 328–341.
46. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Advances in neural information processing systems* **2014**, *27*.
47. Liao, Y.; Xie, J.; Geiger, A. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *Pattern Analysis and Machine Intelligence (PAMI)* **2022**.
48. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572* **2014**.
49. Santara, A.; Rudra, S.; Buridi, S.A.; Kaushik, M.; Naik, A.; Kaul, B.; Ravindran, B. Madras: Multi Agent Driving Simulator. *Journal of Artificial Intelligence Research* **2021**, *70*, 1517–1555.
50. Hu, S.; Liu, J.; Kang, Z. DeepLabV3+/Efficientnet Hybrid Network-Based Scene Area Judgment for the Mars Unmanned Vehicle System. *Sensors* **2021**, *21*, 8136.
51. 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 International Journal of Geo-Information* **2019**, *8*, 390.
52. 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.
53. Chen, Z.; Zhang, J.; Zhang, Y.; Huang, Z. Traffic Accident Data Generation Based on Improved Generative Adversarial Networks. *Sensors* **2021**, *21*, 5767.
54. Zhou, Y.; Fu, R.; Wang, C.; Zhang, R. Modeling Car-Following Behaviors and Driving Styles with Generative Adversarial Imitation Learning. *Sensors* **2020**, *20*, 5034.
55. Lee, J.; Shiotsuka, D.; Nishimori, T.; Nakao, K.; Kamijo, S. Gan-Based Lidar Translation between Sunny and Adverse Weather for Autonomous Driving and Driving Simulation. *Sensors* **2022**, *22*, 5287.
56. Musunuri, Y.R.; Kwon, O.-S.; Kung, S.-Y. SRODNet: Object Detection Network Based on Super Resolution for Autonomous Vehicles. *Remote Sensing* **2022**, *14*, 6270.
57. Choi, W.; Heo, J.; Ahn, C. Development of Road Surface Detection Algorithm Using Cyclegan-Augmented Dataset. *Sensors* **2021**, *21*, 7769.
58. Lee, D. Driving Safety Area Classification for Automated Vehicles Based on Data Augmentation Using Generative Models. *Sustainability* **2024**, *16*, 4337.
59. Sighencea, B.I.; Stanciu, R.I.; Căleanu, C.D. A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction. *Sensors* **2021**, *21*, 7543.
60. 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 Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021); 2021.
61. Waymo - Self-Driving Cars - Autonomous Vehicles - Ride-Hail Available online: https://waymo.com/ (accessed on 17 July 2024).
62. 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, 2019; pp. 1743–1751.
63. 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:1910.03088 [cs, eess]* **2019**.

64. 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, 2020; pp. 687–693.
65. 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 preprint arXiv:1905.01489* **2019**.
66. Pinggera, P.; Ramos, S.; Gehrig, S.; Franke, U.; Rother, C.; Mester, R. Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles. In 2016 IEEE. In Proceedings of the RSJ International Conference on Intelligent Robots and Systems (IROS); pp. 1099–1106.
67. 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); 2019.
68. Blum, H.; Sarlin, P.-E.; Nieto, J.; Siegwart, R.; Cadena, C. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *International Journal of Computer Vision* **2021**, *129*, 3119–3135, doi:10.1007/s11263-021-01511-6.
69. 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); 2019; pp. 2152–2161.
70. 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, 2021; pp. 3095–3101.
71. 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); 2013.
72. Deruyttere, T.; Vandenhende, S.; Grujicic, D.; Van Gool, L.; Moens, M.-F. Talk2Car: Taking Control of Your Self-Driving Car. In Proceedings of the 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, 2019.
73. 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 Robotics and Automation Letters* **2018**, *3*, 2032–2039, doi:10.1109/LRA.2018.2800793.
74. Jeong, J.; Cho, Y.; Shin, Y.-S.; Roh, H.; Kim, A. Complex Urban Dataset with Multi-Level Sensors from Highly Diverse Urban Environments. *International Journal of Robotics Research* **2019**, *38*, 642–657.
75. Zendel, O.; Schörghuber, M.; Rainer, B.; Murschitz, M.; Beleznai, C. Unifying Panoptic Segmentation for Autonomous Driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2022; pp. 21351–21360.
76. Chan, R.; Lis, K.; Uhlemeyer, S.; Blum, H.; Honari, S.; Siegwart, R.; Fua, P.; Salzmann, M.; Rottmann, M. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In Proceedings of the Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks; Vanschoren, J., Yeung, S., Eds.; 2021; Vol. 1.
77. Braun, M.; Krebs, S.; Flohr, F.B.; Gavrila, D.M. EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, 1–1, doi:10.1109/TPAMI.2019.2897684.
78. 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 Proceedings of the Asian Conference on Computer Vision (ACCV); December 2022; pp. 4361–4377.
79. 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 preprint arXiv:2112.02604* **2021**.
80. 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, 2022; pp. 370–387.
81. 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, November 22-25, 2021; BMVA Press, 2021; p. 277.

82. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; pp. 20178–20188.
83. Nekrasov, A.; Zhou, R.; Ackermann, M.; Hermans, A.; Leibe, B.; Rottmann, M. OoDIS: Anomaly Instance Segmentation Benchmark. *arXiv preprint arXiv:2406.11835* **2024**.
84. Belkada, Y.; Bertoni, L.; Caristan, R.; Mordan, T.; Alahi, A. Do Pedestrians Pay Attention? Eye Contact Detection in the Wild. *arXiv preprint arXiv:2112.04212* **2021**.
85. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In Proceedings of the CVPR; 2016.
86. Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang; Tang, X. Spatial As Deep: Spatial CNN for Traffic Scene Understanding. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI); February 2018.
87. 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. *CoRR* **2020**, *abs/2004.06320*.
88. 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 Transactions on Pattern Analysis & Machine Intelligence* **5555**, 1–1, doi:10.1109/TPAMI.2022.3150906.
89. 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 The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR); 2023.
90. Malla, S.; Dariush, B.; Choi, C. TITAN: Future Forecast Using Action Priors. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; pp. 11186–11196.
91. 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 Transactions on Intelligent Transportation Systems* **2018**, *20*, 1633–1643.
92. Bao, W.; Yu, Q.; Kong, Y. Uncertainty-Based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In Proceedings of the ACM Multimedia Conference; May 2020.
93. 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 Proc. International Conference on Robotics and Automation, in press; 2019.
94. 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, 2019; pp. 9711–9717.
95. Pandey, G.; McBride, J.R.; Eustice, R.M. Ford Campus Vision and Lidar Data Set. *The International Journal of Robotics Research* **2011**, *30*, 1543–1552, doi:10.1177/0278364911400640.
96. Lambert, J.; Hays, J. Trust, but Verify: Cross-Modality Fusion for HD Map Change Detection. In Proceedings of the Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021); 2021.
97. Che, Z.; Li, G.; Li, T.; Jiang, B.; Shi, X.; Zhang, X.; Lu, Y.; Wu, G.; Liu, Y.; Ye, J. D ^ 2-City: A Large-Scale Dashcam Video Dataset of Diverse Traffic Scenarios. *arXiv preprint arXiv:1904.01975* **2019**.
98. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024; pp. 4472–4482.
99. Yin, G.; Liu, B.; Zhu, H.; Gong, T.; Yu, N. A Large Scale Urban Surveillance Video Dataset for Multiple-Object Tracking and Behavior Analysis. *CoRR* **2019**, *abs/1904.11784*.
100. Brahmbhatt, S. A Dataset and Model for Crossing Indian Roads. In Proceedings of the Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing; 2022; pp. 1–8.
101. 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 preprint arXiv:2109.07648* **2021**.

102. 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 Proceedings of the British Machine Vision Conference (BMVC); 2019.
103. 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, 2023; pp. 137–154.
104. Choi, M.; Goel, H.; Omama, M.; Yang, Y.; Shah, S.; Chinchali, S. Towards Neuro-Symbolic Video Understanding. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV); September 2024.
105. Oliveira, I.O. de; 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, doi:10.1109/ACCESS.2021.3097964.
106. 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., February 2021.
107. Sivaraman, S.; Trivedi, M.M. A General Active-Learning Framework for On-Road Vehicle Recognition and Tracking. *IEEE Transactions on Intelligent Transportation Systems* **2010**, *11*, 267–276, doi:10.1109/TITS.2010.2040177.
108. 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, 2018; pp. 621–635.
109. 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 Robotics and Automation Letters* **2022**, *7*, 10914–10921.
110. 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, 2020; pp. 4023–4029.
111. Falkner, J.K.; Schmidt-Thieme, L. Learning to Solve Vehicle Routing Problems with Time Windows through Joint Attention. *arXiv preprint arXiv:2006.09100* **2020**.
112. 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. In Proceedings of the Transactions on Machine Learning Research; 2023.
113. Hu, H.-N.; Yang, Y.-H.; Fischer, T.; Darrell, T.; Yu, F.; Sun, M. Monocular Quasi-Dense 3d Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**.
114. 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, November 21-24, 2022; BMVA Press, 2022.
115. 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, December 2022; pp. 4800–4822.
116. 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, 2020; Vol. 84287, p. V002T36A011.
117. Deshpande, A.M.; Minai, A.A.; Kumar, M. Robust Deep Reinforcement Learning for Quadcopter Control. *IFAC-PapersOnLine* **2021**, *54*, 90–95, doi:https://doi.org/10.1016/j.ifacol.2021.11.158.
118. Bhattacharyya, M.; Nag, S.; Ghosh, U. Deciphering Environmental Air Pollution with Large Scale City Data. In Proceedings of the Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22; International Joint Conferences on Artificial Intelligence Organization, 2022.
119. 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); 2021; pp. 934–939.

120. 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.
121. Massimiliano, V. Semantic Segmentation on Cityscapes Using Segmentation Models Pytorch.
122. 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 preprint arXiv:2204.02681* **2022**.
123. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Road Scenes. *arXiv preprint arXiv:2101.06085* **2021**.
124. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 9716–9725.
125. 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; 2019; pp. 3552–3561.
126. Nirkin, Y.; Wolf, L.; Hassner, T. Hyperseg: Patch-Wise Hypernetwork for Real-Time Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 4061–4070.
127. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019; pp. 12607–12616.
128. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *International journal of computer vision* **2021**, *129*, 3051–3068.
129. Hu, P.; Caba, F.; Wang, O.; Lin, Z.; Sclaroff, S.; Perazzi, F. Temporally Distributed Networks for Fast Video Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; pp. 8818–8827.
130. 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, 2020; pp. 933–938.
131. 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 preprint arXiv:2303.17895* **2023**.
132. 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, 2023; pp. 2774–2781.
133. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; pp. 8394–8405.
134. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; pp. 6792–6802.
135. 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 preprint arXiv:2211.13529* **2022**.
136. 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 Proceedings of the AAAI Conference on Artificial Intelligence; 2023; Vol. 37, pp. 1179–1187.
137. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3d Object Detection and Tracking. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 11784–11793.
138. 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, August 23–28, 2020, Proceedings, Part XXV 16; Springer, 2020; pp. 581–597.
139. Chen, D.; Krähenbühl, P. Learning from All Vehicles. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; pp. 17222–17231.
140. Chitta, K.; Prakash, A.; Jaeger, B.; Yu, Z.; Renz, K.; Geiger, A. Transfuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 12878–12895.

 doi:10.20944/preprints202407.1857.v1

141. 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 preprint arXiv:2406.10165* **2024**.
142. Jaeger, B.; Chitta, K.; Geiger, A. Hidden Biases of End-to-End Driving Models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; pp. 8240–8249.