Recent Advancements in End-to-End Autonomous Driving Using Deep Learning: A Survey

Pranav Singh Chib and Pravendra Singh

***Abstract*—End-to-End driving is a promising paradigm as it circumvents the drawbacks associated with modular systems, such as their overwhelming complexity and propensity for error propagation. Autonomous driving transcends conventional traffic patterns by proactively recognizing critical events in advance, ensuring passengers safety and providing them with comfortable transportation, particularly in highly stochastic and variable traffic settings. This article presents a comprehensive review of the End-to-End autonomous driving stack. It provides a taxonomy of automated driving tasks wherein neural networks have been employed in an End-to-End manner, encompassing the entire driving process from perception to control. Recent developments in End-to-End autonomous driving are analyzed, and research is categorized based on underlying principles, methodologies, and core functionality. These categories encompass sensorial input, main and auxiliary output, learning approaches ranging from imitation to reinforcement learning, and model evaluation techniques. The survey incorporates a detailed discussion of the explainability and safety aspects. Furthermore, it assesses the state-of-the-art, identifies challenges, and explores future possibilities.**

***Index Terms*—Autonomous driving, end-to-end driving, intelligent transportation system, deep learning.**

I. Introduction

AUTONOMOUS driving refers to the capability of a vehicle to drive partly or entirely without human intervention. The modular architecture [1], [2], [3], [4], [5] is a widely used approach in autonomous driving systems, which divides the driving pipeline into discrete sub-tasks. This architecture relies on individual sensors and algorithms to process data and generate control outputs. It encompasses interconnected modules, including perception, planning, and control. However, the modular architecture has certain drawbacks that impede further advancements in autonomous driving (AD). One significant limitation is its susceptibility to error propagation. For instance, errors in the perception module of a self-driving vehicle, such as misclassification, can propagate to subsequent planning and control modules, potentially leading to unsafe behaviors. Additionally, the complexity of managing interconnected modules and the computational inefficiency of processing data at each stage pose additional challenges associated with the modular approach. To address these shortcomings, an alternative approach called End-to-End driving [6], [7], [8], [9], [10], [11] has emerged. This approach aims to overcome the limitations of the modular architecture.

The End-to-End approach streamlines the system, improving efficiency and robustness by directly mapping sensory input to control outputs. The benefits of End-to-End autonomous driving have garnered significant attention in the research community. Firstly, End-to-End driving addresses the issue of error propagation, as it involves a single learning task pipeline [12], [13] that learns task-specific features, thereby reducing the likelihood of error propagation. Secondly, End-to-End driving offers computational advantages. Modular pipelines often entail redundant computations, as each module is trained for task-specific outputs [4], [5]. This results in unnecessary and prolonged computation. In contrast, End-to-End driving focuses on the specific task of generating the control signal, reducing the need for unnecessary computations and streamlining the overall process. End-to-End models were previously regarded as "black boxes", lacking transparency. However, recent methodologies have improved interpretability in End-to-End models by generating auxiliary outputs [7], [13], attention maps [9], [14], [15], [16], [17], [18], and interpretable maps [8], [12], [18], [19], [20], [21]. This enhanced interpretability provides insights into the root causes of errors and model decision-making. Furthermore, End-to-End driving demonstrates resilience to adversarial attacks. Adversarial attacks [22] involve manipulating sensor inputs to deceive or confuse autonomous driving systems. In End-to-End models, it is challenging to identify and manipulate the specific driving behavior triggers as it is unknown what causes specific driving patterns. Lastly, End-to-End driving offers ease of training. Modular pipelines require separate training and optimization of each task-driven module, necessitating domain-specific knowledge and expertise. In contrast, End-to-End models can learn relevant features and patterns [23], [24] directly from raw sensor data, reducing the need for extensive engineering and expertise.

Related Surveys: A number of related surveys are available, though their emphasis differs from ours. The author Yurtsever et al. [25] covers the autonomous driving domain with a primary emphasis on the modular methodology. Several past surveys center around specific learning techniques, such as imitation learning [26] and reinforcement learning [27]. A few end-to-end surveys, including the work by Tampuu et al. [28], provide an architectural overview of the complete end-to-end driving pipeline. Recently, Chen et al. [29] discuss the methodology and

Manuscript received 4 September 2023; accepted 18 September 2023. Date of publication 22 September 2023; date of current version 23 February 2024. *(Corresponding author: Pravendra Singh.)*

The authors are with the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India (e-mail: pranavs_chib@cs.iitr.ac.in; pravendra.singh@cs.iitr.ac.in).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2023.3318070.

Digital Object Identifier 10.1109/TIV.2023.3318070

challenges in end-to-end autonomous driving in their survey. Our focus, however, is on the latest advancements, including modalities, learning principles, safety, explainability, and evaluation (see Table I).

Motivation and Contributions: The End-to-End architectures have significantly enhanced autonomous driving systems. As elaborated earlier, these architectures have overcome the limitations of modular approaches. Motivated by these developments, we present a survey on recent advancements in End-to-End autonomous driving. The key contributions of this article are threefold. First, this survey exclusively explores End-to-End autonomous driving using deep learning. We provide a comprehensive analysis of the underlying principles, methodologies, and functionality, delving into the latest state-of-the-art advancements in this domain. Second, we present a detailed investigation in terms of modality, learning, safety, explainability, and results, and provide a quantitative summary in Table I. Third, we present an evaluation framework based on both open and closed-loop assessments and compile a summarized list of available datasets and simulators.

Paper Organization: The article is organized as follows. Section II provides an overview of the End-to-End autonomous driving pipeline architecture. This is followed by Sections III and IV, which discuss the input and output modalities of the system. Section V comprehensively covers learning methods, from imitation learning to reinforcement learning. The domain adaptation is explained in Section VI. Next, we explore the safety aspect of End-to-End approaches in Section VII. The importance of explainability and interpretability is discussed in Section VIII. The evaluation of the End-to-End system consists of open and closed-loop evaluations, which are discussed in Section IX. The relevant datasets and the simulator are presented in Section X. Finally, Sections XI and XII provide the future research direction and conclusion, respectively.

II. End-to-End System Architecture

In general, modular systems are referred to as the mediated paradigm and are constructed as a pipeline of discrete components (Fig. 1) that connect sensory inputs and motor outputs. The core processes of a modular system include perception, localization, mapping, planning, and vehicle control [1]. The modular pipeline starts by inputting raw sensory data to the perception module for obstacle detection [30] and localization via the localization module [3]. This is followed by planning and prediction [31] to determine the optimal and safe trajectory for the vehicle. Finally, the motor controller generates commands for safe maneuvering.

On the other hand, direct perception or End-to-End driving directly generates ego-motion from the sensory input. It optimizes the driving pipeline (Fig. 1) by bypassing the sub-tasks related to perception and planning, allowing for continuous learning to sense and act, similar to humans. The first attempt at End-to-End driving was made by Pomerleau Alvinn [32], which trained a 3-layer sensorimotor fully connected network to output the car's direction. End-to-End driving generates ego-motion based on sensory input, which can be of various modalities. However, the prominent ones are the camera [33], [34], [35], Light Detection and Ranging (LiDAR) [6], [7], [10], navigation commands [23], [34], [36], and vehicle dynamics, such as speed [35], [37], [38]. This sensory information is utilized as the input to the backbone model, which is responsible for generating control signals. Ego-motion can involve different types of motions, such as acceleration, turning, steering, and pedaling. Additionally, many models also output additional information, such as a cost map for safe maneuvers, interpretable outputs, and other auxiliary outputs.

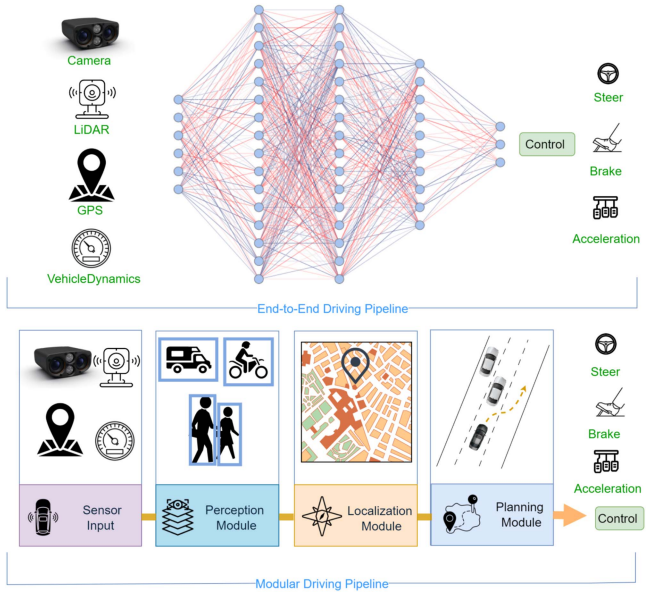


Fig. 1. Comparison between end-to-end and modular pipelines. End-to-End is a single pipeline that generates the control signal directly from perception input, whereas a modular pipeline consists of various sub-modules, each with task-specific functionalities.

There are two main approaches for End-to-End driving: either the driving model is explored and improved via Reinforcement Learning (RL) [21], [38], [39], [40], [41], [42], or it is trained in a supervised manner using Imitation Learning (IL) [6], [7], [15], [17], [18], [19], [43] to resemble human driving behavior. The supervised learning paradigm aims to learn the driving style from expert demonstrations, which serve as training examples for the model. However, expanding an autonomous driving system based on IL [23] is challenging since it is impossible to cover every instance during the learning phase. On the other hand, RL works by maximizing cumulative rewards [40], [44] over time through interaction with the environment, and the network makes driving decisions to obtain rewards or penalties based on its actions. While RL model training occurs online and allows exploration of the environment during training, it is less effective in utilizing data compared to imitation learning. Table I summarizes recent methods in End-to-End driving.

III. Input Modalities

A. Camera

Camera-based methods [6], [9], [12], [13], [14], [23], [34], [35], [38], [45] have shown promising results in End-to-End driving. For instance, Toromanoff et al. [38] demonstrated their

TABLE I
RECENT METHODS IN END-TO-END AUTONOMOUS DRIVING

Paper, Year	Environment	Input modality	Output modality	Learning	Evaluation	Explaibility	Safety	Results	Data
PAD [9], 2023	NuScenes	6 Multi-camera images	Segmentation map, agent future trajectories, future occupancy	Multi-task learning.	min ADE, min FDE, IoU, L2 error	Exp. Via attention mask in planner module	Lowest collision rate, lowest L2 error, Safety boost via goal planner.	L2: 1.65 CR: 0.71	Perception training: 6 epochs Joint training: 20 epochs
ReasonNet [16], 2023	CARLA	Vehicle measurements navigation, LiDAR, RGB	Steering, speed, BEV, brake	Multi-task learning	DOS and LeaderBoard	Attention map	-	DS: 79.95 RC: 89.89 IS: 0.89	Train: 2M frames (eight Towns) Test: Town05
Policy Pre-Training [66], 2023	NuScenes, CARLA	Camera and intrinsics	Steering, throttle	Imitation learning	Longest6, collision rate	Attention activation map	Handles cases where the agent needs to stop	DS: 47.4±5.6 RC: 65.05±5.1 IS: 0.79±0.08 L2: 3.01, CR: 0.94	Train: 40K (Town01, 03, 04, 06) Test: 50 routes
Think Twice [19], 2023	CARLA	RGB camera, LiDAR	BEV feature map, agent future trajectories, control actions	Open-loop imitation learning	Longest6	Explainable via expert BEV feature map	Safety-critical regions, anticipate future motion to avoid collisions	DS: 70.9±3.4 RC: 95.5±2.6 IS: 0.75±0.05	Training: 189K data on four Towns, Roach based teaching
Scaling Vision-based [17], 2023	CARLA	Three cameras (views)	Ego-vehicle's steering angle and acceleration	Self-supervised imitation learning	NoCrash, Leaderboard	Attention map	Pedestrians caused strong braking (0.794), despite green light	DS: 98±1.7 RC: 68±2.7	Train: 15 hours (540K frames) Test: 25 hours (900k frames)
Coaching a Teachable [43], 2023	CARLA	Navigation, 3 RGB cameras	Waypoints, control commands	Knowledge distillation, imitation learning	Longest6, ADE, FDE	-	Attend to safety-critical entities	DS: 73.30±1.07 RC: 87.44±0.28 ADE/FDE: 0.41/0.36	Train: Town01 - Town06
Hidden Biases of [56], 2023	CARLA	Speed, camera and LiDAR	Steer, throttle, and brake, waypoints, path	Imitation learning	Longest6	-	Larger safety distance, safety area	DS: 72 ±3 RC: 95 ± 2	Train: variable size 185k and 555k
KING [6], 2022	CARLA	Front-facing camera and LiDAR	Throttle and steering	Behaviour cloning	Closed-loop (CR), (RC), (IS), (DS)	-	Safety-critical driving scenarios and stress-testing.	DS: 90.20±0.00 RC: 94.42±0.36 IS: 0.96±0.36	Train: 4 GPU hours 80 routes on Town 3,4,5,6.
LAV [10], 2022	CARLA	Front-facing camera, LiDAR	Single steering and acceleration command	Knowledge distillation, imitation learning	Longest6 (DS), (RC),(IS), (CR), (PC), (LC), (RV)	Spatial feature 2D map	Scenario such as pedestrians crossing, lane change are modeled	DS: 61.85 RC: 94.46 IS: 0.64	Train: Four Towns, 186K frames Test: Town02 and Town05
TransFuse [7], 2022	CARLA	RGB, LiDAR	Waypoints, steer, throttle, and brake	Behaviour cloning	Longest 6 (DS), (RC), (IPK), (IS)	Explainable via auxiliary output like depth, semantic, HD map, planner A*	Global safety heuristic	DS: 61.18 RC: 86.69 IS: 0.71	Train: 2500 routes on eight Towns, 228k frames Test: 36 route
Learning to Drive [11], 2022	NuScenes, youtube	Front-facing camera	Steering, throttle, brake, and velocity	Conditional behavior cloning	CARLA benchmark, F1 metric	-	-	F1 : 75.0 7 perc. increase in RC	Train: 120 h YouTube, Test: Town01 and 02 with 40K transition
HACO [40], 2022	CARLA	Current state, navigation information	Acceleration, brake and steering	Reinforcement learning	Safety violation data usage, success rate	-	Safety violation cost the episode	SV: 11.84 SR: 0.35	Train: 50 min HACO training. 35k transition
PlanT [18], 2022	CARLA	Camera , LiDAR, route, object	Waypoints, predicting the future attributes of other vehicles	Imitation learning	Longest6 CARLA (DS),(RC), (IS), (IPK), (IT)	Post hoc explanations, attention weights for identify relevant objects.	Identification of collision free routes, RFDS algorithm for collision avoidance	DS: 81.36±6.54 RC: 93.55±2.62 IS: 0.87±0.05	Train: 3.2 hours on PlanT, 95 hours of driving.
Trajectory-guided [13], 2022	CARLA	Monocular camera	Steering, throttle and brake	Behaviour cloning	CARLA (DS), (RC), (IS)	Auxiliary tasks include value and speed head	Reduce collision via long prediction, less infraction, control model preform good	DS: 75.14 RC: 85.63 IS: 0.87	Train: Town01, 03, 04, 06 Test: Town02, 05
Safety- Enhanced Autonomous Driving [8], 2022	CARLA	3 RGB, 1 LiDAR	10 Waypoints, steering, acceleration	InterFuser	CARLA leaderboard, CARLA 42 Routes benchmark, (RC),(IS),(DS)	Intermediate interpretable features output from transformer decoder (safety map, object density)	Safe set contain only safe actions, safety sensitive output via InterFuser	DS: 76.18 RC: 88.23 IS: 0.84	Train: 3M frames or 410 hours (eight Towns)
ST-P3 [20], 2022	NuScenes, CARLA	6 Camera(NuScenes), 4 cameras (CARLA), navigation command	Steering, throttle and brake	ST-P3	Open loop: IOU, PQ, RQ, SQ, L2 error Closed: (DS), (RC)	Interpretable map lanes and area which is drivable	Safety Cost function for the jerk action penalize	DS: 55.14 RC: 86.74 L2: 2.90, CR: 1.27	Train: 26124 samples Validation: 5719 samples
Safe Driving via Expert Guided Policy [67], 2022	MetaDrive	Camera	Control signal	Expert-in-the-loop reinforcement learning	MetaDriv benchmark	-	Guardian to ensure training safety	ER: 388.37 ±10.01 EC: 0.56 ±0.35 SR: 0.85 ±0.05	Train: 100 scenes Test: 50 scenes
COOPERNAUT [68] , 2022	CARLA	Front-facing camera, LiDAR	Control signal	Behaviour cloning	AUTOCASTSIM	-	Reduces safety hazards for line-of-sight sensing.	SR: 90.3±1.2 CR: 4.5±3.1	Train: 12 traces + 84 traces Test : 27 accident-prone traces
Human-AI Shared Control via Policy [21], 2022	MetaDrive	Current state, goal state	Control signal	Reinforcement learning	MetaDrive and Pybullet-A1	Interpretable control interface	Human-AI, safety guarantee 95prec. success rate	EC: 0.05 ±0.08 SR: 0.95 ± 0.02	Train: 50 training scenario Test: 20 test scene
MMFN: Multi-Modal Fusion-Net for [51], 2022	CARLA	HD map and radar on top of the LiDAR and camera	Steering, throttle and brake	Imitation learning	CARLA LeaderBoard	-	Expert has more awareness of safe driving	DS: 22.8 RC: 47.22	Train: 207K frames Test: 20 routes
CADRE: A Cascade Deep Reinforcement [42], 2022	CARLA	Front-view camera, position, orientation and speed	Control signal	Imitation learning	CARLA NoCrash	-	-	VA: 81/81 PA: 76/78	Train:25 training routes Test: Town02
Model-Based Imitation Learning for [69], 2022	CARLA	RGB image, route	Vehicle control, BEV Segmentation	Model-based imitation learning	CARLA LeaderBoard	BEV semantic segmentation for interpretability	-	DS: 61.1 ± 3.2 RC: 97.4± 0.8 IS: 63.0 ± 3.0	Test: Town05, routes Train: Four different training towns total of 2.9M frames.
LookOut: Diverse Multi-Future Prediction and Planning [65], 2021	ATG4D, Lidarsim	LiDAR, navigation.	Trajectory, vehicle control	Cost learning	Open loop: mAP, mSADE, mSADE, PlanASD Cloose loop: Lidarsim	-	Cost function include driving including safety, comfort, traffic-rules.	CR: 7.93 Progress:62.65 Jerk: 4.69	Train: one million frames Test: Lidarsim
MP3: A Unified Model to Map, Perceive, Predict and Plan [70], 2021	URBANEXPERT	Raw sensor data and a high-level command	Control command, dynamic occupancy field	MP3 -	Closed-loop: Lidarsim Open loop: L2	Interpretable cost functions, dynamic occupancy field, Interpretable Scene repr.	Penalize trajectories where the SDV overlaps occupied regions, penalize jerk, lateral acceleration	L2: 12.95 Collision: 1037.08 Jerk: 1.64 Success Pre: 74.39	Train: 5000 scenarios Test: 1000
Object-Aware Regularization for Addressing Causal [71], 2021	CARLA	RGB image	Control signal	Behaviour cloning	Atari environment, CARLA .	-	Policy safe adaptation.	Straight: 87.4±4.4 turn: 70.0± 7.2 Nav: 35.7±10.2	Train: 150 demonstrations Test : 25 routes
GRI: General Reinforced Imitation and its [72], 2021	CARLA	3 RGB camera view	Control signal	Off-policy reinforcement learning	CARLA Leaderboard, NoCrash, Mujoco benchmark	-	-	DS: 36.79 RC: 61.85 IS: 0.60	Train: 60M steps Test: with 12M and 16M steps
Multi-Modal Fusion Transformer [67], 2021	CARLA	Front-facing camera and LiDAR	4 Waypoints, steer, throttle, and brake	Imitation learning	CARLA LeaderBoard	Attention map visualizations	Handle adversarial scenarios in urban driving, e.g., hard turnings.	DS: 33.15 ± 4.04 RC: 56.36 ± 7.14	Train: 7 towns Test: Twon05
Learning by Watching [36], 2021	CARLA	Speed, high-level navigation	Waypoints, steer, throttle, and brake	Imitation learning	CARLA NoCrash benchmarks	BEV visibility map	Avoid an unsafe maneuver.	NC-R: 92 NC-D: 24 OR: 92	Train: Town 1 Test: Town 2
NEAT [12], 2021	CARLA	RGB cameras, and intrinsics, locations,speed	Waypoints, BEV as a auxiliary output	Behaviour cloning	CARLA Leaderboard. (RC), (IS), (DS)	NEAT intermediate representations provides interpretable attention map	At that time highest safety among other methods on the CARLA.	DS: 24.08±3.30 RC: 59.94±0.50 IS: 0.49±0.02	Train: 8 towns Test: (Town01-Town 06) 100 secret routes
End-to-End Urban Driving [39], 2021	CARLA	Wide-angle camera image with a 100° horizontal FOV	Steering, throttle and brake	Reinforcement learning expert, imitation learning	CARLA NoCrash and LeaderBoard	-	-	DS: 55.27±1.43 RC: 88.16±1.52 IS: 0.62±0.02	Off-policy dataset 80 episodes, Train: 50 routes Test: 26 routes
Learning to drive from [37], 2021	CARLA	RGB images and speed readings	Steering, throttle and brake	Reinforcement learning, policy distillation	CARLA LeaderBoard	-	Action-values based on the current ego-vehicle state	DS: 17.36±2.95 RC: 43.46±2.99 IS: 0.54±0.06	Train: 69 hours about 1m frames Test: 270K frames
Safe Local Motion Planning with Self-Supervised [48], 2021	NuScenes, CARLA	LiDAR	Control signal, BEV	Behaviour cloning	CARLA NoCrash	Object-centric representation	Safe planner, maintaining a wide safety margin, avoid safety critical situations	(Success rate) NC-E:66 ± 3 NC-R: 73± 1 NC-D: 44 ± 5	Train: Town 1 Test: Town 2 Train: 850 scenes(NuScenes) Test: 150 scenes
Carl-Lead: Lidar-based End-to-End [49], 2021	CARLA	LiDAR	Steering, throttle, brake, HD map	Reinforcement learning	CARLA NoCrash	Saliency maps for visualizing model predictions	Collision reward punishment for unsafe driving	(Success rate) NC-R: 93.50 NC-D: 93.00	Train: 677.7K interaction steps Test: 14,400 episodes
A Versatile and Efficient Reinforcement Learning [73], 2021	CARLA, BDD100k	RGB image	Control signal	Reinforcement learning, imitation learning	NoGap on CARLA	Interpretable segmentation map	Move the vehicle to safe state	MPI (m): Turn:7.2, Straight : 4.1 SR: Turn: 53.2 Straight: 16.7 RL: 332.6, IL:180.9	Train: 28h of driving and 2.5M RL steps
Multi-task Learning with Attention for End-to-end [46], 2021	CARLA	Monocular RGB, velocity	Steering, throttle and brake	Conditional imitation learning, multitask learning	CARLA NoCrash , CoRL2017 benchmark	-	-	Straight :99 ± 1 One Turn: 99 ± 1 NC-E: 81± 11 NC-R :67 ± 9 NC-D:23 ± 5	Train: Town01 (466,000 frames) Test: Town02
End-to-End Model-Free Reinforcement [38], 2020	CARLA	Front-facing camera , speed	5 steer actions, 3 values for throttle, one for brake	Reinforcement learning	CARLA NoCrash, LeaderBoard	Coin implicit affordances	-	NC-R: 96 NC-D:70 NC-E: 100	Train: 20M iterations Town05 Test: Town02
Learning by cheating [33], 2020	CARLA	Front-facing camera , speed, navigation command	Steering, throttle and brake	On-policy imitation learning	CARLA NoCrash and LeaderBoard	Map representation	Conduct a separate infraction analysis	NC-E: 100 NC-R: 94±4 NC-D: 85±1	DAgger training. Train: 157K frames, 4 hours driving Town1, Test: Town2
Learning Situational Driving [34], 2020	CARLA	Front-facing camera , speed, navigation command	Longitudinal and lateral control values	Behaviour cloning	CARLA NoCrash, LeaderBoard	Situation-specific predictions can be inspected at test time	Learned agent to adhere to traffic rules and safety	NC-R: 64 NC-D: 32	Train: Town1 Test: Town2
SAM [35], 2020	CARLA	Image, self-speed, turning command	Brake, gas, and steering angle	Conditional imitation learning	Traffic-school benchmark, CARLA NoCrash	-	Stop intentions help avoid hazardous traffic situations	NC-E: 83±1 NC-R: 68±7 NC-D: 29±2	Train : 10 hours (360K frames) Town01.
Multi-task Learning with Future States for Vision [74], 2020	CARLA	Three RGB cameras	Control signal	Multi-task learning, conditional imitation learning	NoCrash, AnyWeather	-	Localization tasks is useful for safe driving.	NC-E: 92, NC-R 66, NC-D: 32,SR: 93.2	Train: 100 hours
DSDNet [50], 2020	NuScenes, ATG4D, CARLA	LiDAR, HD map	Steering, speed	Imitation learning	CR, L2, CARLA	Learn interpretable intermediate results	Safer planning by cost function	L2: 1.22,Lane vio: 1.55 IOU: 55.4 Min MSD: 0.213	Train: 60K, 1000, 5000 samples Test: 17K, 500 samples
Perceive, Predict, and Plan: Safe Motion [75], 2020	Real scenario	Raw sensor data, HD map, high level route	Control action	Imitation learning, inverse RL	L2, CR, jerk and lateral acceleration	Interpretable Semantic Representations(occupancy)	Planner learn safety cost, safety buffer	CR: 1.78 L2:3.34 Jerk: 1.27 Lat. acc: 2.89	Train: 6100 scenarios Test: 1500 scenarios.
Urban driving with condition [24], 2020	Real scenario	Camera , navigation command	Steering, speed	Imitation learning	Successful turns, CR,TV	Using pre trained perception	Safety-driver in loop	MAE: 0.0715	Train: 30h Test: 26 routes

Route Completion (RC), Infraction Score/penalty (IS), Driving score (DS), Collisions pedestrians (CP)/(PC), Collisions vehicles (CV), Collisions layout (CL)/(LC), Red light infractions (RLI), Red light violation (RV), Stop sign infractions (SSI), Off-coad infractions (OI), Route deviations (RD), Agent blocked (AB), Average Displacement Error (ADE), Final Displacement Error (FDE), Intersection over Union (IOU), Panoptic Quality (PQ), Recognition Quality (RQ), Segmentation Quality (SQ), Instance Contrastive Pair (ICP), Action Contrastive Pair (ACP), Driving in Occlusion Simulation (DOS), Episodic Cost (EC), Success Rate (SR), Collision Rate (CR), Safety Violation (SV), Success Rate (SR), Episodic Return (ER), Episodic Cost (EC), Vehicle Avoidance (VA), Pedestrian Avoidance (PA), Meters Per Intervention(MPI), Reinforcement Learning (RL), Imitation Learning (IL), NoCrash Regular Traffic (NC-R), NoCrash Dense Traffic (NC-D), NoCrash Empty (NC-E).

capabilities by winning the CARLA 2019 autonomous driving challenge using vision-based approaches in an urban context. The use of monocular [11], [13], [38], [46] and stereo vision [15], [17], [43] camera views is a natural input modality for image-to-control End-to-End driving. Xiao et al. [47] employed inputs consisting of a monocular RGB image from a forward-facing camera and the vehicle speed. Wu et al. [15], Xiao et al. [17], Zhang et al. [43] utilize camera-only modality to generate high-level instructions for lane following, turning, stopping and going straight using imitation learning.

B. LiDAR

Another significant input source in self-driving is the Light Detection and Ranging (LiDAR) sensor. LiDAR [20], [48], [49], [50] is resistant to lighting conditions and offers accurate distance estimates. Compared to other perception sensors, LiDAR data is the richest and provides the most comprehensive spatial information. It utilizes laser light to detect distances and generates PointClouds, which are 3D representations of space where each point includes the (x, y, z) coordinates of the surface that reflected the sensor's laser beam. When localizing a vehicle, generating odometry measurements is critical. Many techniques utilize LiDAR for feature mapping in Birds Eye View (BEV) [9], [16], [19], High Definition (HD) map [20], [51], and Simultaneous Localization and Mapping (SLAM) [52]. Shenoi et al. [53] have shown that adding depth and semantics via LiDAR has the potential to enhance driving performance. Liang et al. [54], [55] utilized point flow to learn the driving policy in an End-to-End manner.

C. Multi-Modal

Multimodality [8], [10], [16], [18], [56] outperforms single modality in crucial perception tasks and is particularly well-suited for autonomous driving applications, as it combines multi-sensor data. There are three broad categorizations for utilizing information depending on when to combine multi-sensor information. In early fusion, sensor data is combined before feeding them into the learnable End-to-End system. Chen et al. [10] and Xiao et al. [47] use a network that accepts (RGB + Depth) channel inputs. The network modifies just the first convolutional layer to account for the additional input channel, while the remaining network remains unchanged. Renz et al. [18] fuse object-level input representation using a transformer encoder. The author combinedly represents a set of objects as vehicles and segments of the routes.

In mid-fusion, information fusion is done either after some preprocessing stages or after some feature extraction. Zhou et al. [57] perform information fusion at the mid-level by leveraging the complementary information provided by both the bird's-eye view (BEV) and perspective views of the LiDAR point cloud. Transfuser [7] addresses the integration of image and LiDAR modalities using self-attention layers. They utilized multiple transformer modules at multiple resolutions to fuse intermediate features. Obtained feature vector forms a concise representation which an MLP then processes before passing to an auto-regressive waypoint prediction network. In late fusion, inputs are processed separately, and their output is fused and further processed by another layer. Some authors [47], [54], [55], [58] use a late fusion architecture for LiDAR and visual modalities, in which each input stream is encoded separately and concatenated.

D. Navigational Inputs

End-to-End navigation input can originate from the route planner [8], [14], [59] and navigation commands [33], [60], [61], [62]. Routes are defined by a sequence of discrete endpoint locations in Global Positioning System (GPS) coordinates provided by a global planner [14]. The TCP model [13] is provided with correlated navigation directives like lane keeping, left/right turns, and the destination. This information is used to produce the control actions.

Shao et al. [8] propose a technique that guides driving using these sparse destination locations instead of explicitly defining discrete navigation directives. PlanT [18] utilizes point-to-point navigation based on the input of the goal location. FlowDriveNet [63] considers both the global planner's discrete navigation command and the coordinates of the navigation target. Hubschneider et al. [60] include a turn indicator command in the driving model, while Codevilla et al. [61] utilize a CNN block for specific navigation tasks and a second block for subset navigation. In addition to the aforementioned inputs, End-to-End models also incorporate vehicle dynamics, such as ego-vehicle speed [12], [34], [37], [38].

IV. Output Modalities

A. Waypoints

Predicting future waypoints is a higher-level output modality. Several authors [6], [7], [10], [64] use an auto-regressive waypoint network to predict differential waypoints. Trajectories [8], [13], [19], [59], [65] can also represent sequences of waypoints in the coordinate frame. The network's output waypoints are converted into low-level steering and acceleration using Model Predictive Control (MPC) [4] and Proportional Integral Derivative (PID) [5]. The longitudinal controller considers the magnitude of a weighted average of vectors between successive time-step waypoints, while the lateral controller considers their direction. The ideal waypoint [38] relies on desired speed, position, and rotation. The lateral distance and angle must be minimized to maximize the reward (or minimize the deviation). The benefit of utilizing waypoints as an output is that they are not affected by vehicle geometry. Additionally, waypoints are easier to analyze by the controller for control commands such as steering. Waypoints in continuous form can be transformed into a specific trajectory. Zhang et al. [39] and Zhou et al. [59] utilize a motion planner to generate a series of waypoints that describe the future trajectory. LAV [10] predicts multi-modal future trajectories for all detected vehicles, including the ego-vehicle. They use future waypoints to represent the motion plan.

B. Cost Function

Many trajectories and waypoints are possible for the safe maneuvering of the vehicle. The cost [20], [21], [41], [65], [69], [75] is used to select the optimal one among the possibilities. It assigns a weight (positive or negative score) to each trajectory based on parameters defined by the end user, such as safety, distance traveled, comfort, and others. Rhinehart et al. [76] and Chen et al. [10] refine control using the predictive consistency map, which updates knowledge at test time. They also evaluate the trajectory using an ensemble expert likelihood model. Prakash et al. [14] utilize object-level representations to analyze collision-free routes. Zeng et al. [75] employ a neural motion planner that uses a cost volume to predict future trajectories. Hu et al. [20] employ a cost function that takes advantage of the learned occupancy probability field, represented by segmentation maps, and prior knowledge such as traffic rules to select the trajectory with the minimum cost. Regarding safety cost functions, Zhao et al. [35], Chen et al. [77], and Shao et al. [8] employ safety maps. They analyze actions within the safe set to create causal insights regarding hazardous driving situations.

C. Direct Control and Acceleration

Most of the End-to-End models [23], [33], [38], [59], [60], [61], [62], [76] provide the steering angle and speed as outputs at a specific timestamp. The output control needs to be calibrated based on the vehicle's dynamics, determining the appropriate steering angle for turning and the necessary braking for stopping at a measurable distance.

D. Auxiliary Output

The auxiliary output can provide additional information for the model's operation and the determination of driving actions. Several types of auxiliary outputs include the segmentation map [7], [9], BEV map [9], [12], [16], [19], future occupancy [9], [18], [69], [75] of the vehicle, and interpretable feature map [7], [8], [12], [18], [20], [75]. These outputs provide additional functionality to the End-to-End pipeline and help the model learn better representations. The auxiliary output also facilitates the explanation of the model's behavior [7], [69], as one can comprehend the information and infer the reasons behind the model's decisions.

V. Learning Approaches

The following sections discuss various learning approaches in End-to-End Driving, including imitation learning and reinforcement learning.

A. Imitation Learning

Imitation learning (IL) [10], [13], [14], [18], [73] is based on the principle of learning from expert demonstrations. These demonstrations train the system to mimic the expert's behavior in various driving scenarios. Large-scale expert driving datasets are readily available, which can be leveraged by imitation learning [47] to train models that perform at human-like standards.

The main objective is to train a policy $\pi_\theta(s)$ that maps each given state to a corresponding action (Fig. 2) as closely as possible to the given expert policy π^*, given an expert dataset with state action pair (s, a):

$$\arg\min_{\theta} E_{s\ \sim P(s|\theta)}\ L(\pi^*(s), \pi_\theta(s)) \tag{1}$$

where $P(s \mid \theta)$ represents the state distribution of the trained policy π_θ.

Behavioural Cloning (BC), Direct Policy Learning (DPL), and Inverse Reinforcement Learning (IRL) are extensions of imitation learning in the domain of autonomous driving.

1) Behavioural Cloning: Behavioural cloning [7], [12], [13], [23], [34], [67], [78] is the supervised imitation learning task where the goal is to treat each state-action combination in expert distribution as an Independent and Identically Distributed (I.I.D) example and minimise imitation loss for the trained policy:

$$\arg\min_{\theta} E_{(s,a^*)\ \sim P^*}\ L(a^*, \pi_\theta(s)) \tag{2}$$

where $p^*(s \mid \pi^*)$ is an expert policy state distribution, and (state s, action a^*) is provided by expert policy p^*.

Prakash et al. [14], Chitta et al. [7], NEAT [12], Ohn et al. [34] utilize a policy that maps input frames to low-level control signals in terms of waypoints. These waypoints are then fed into a PID to obtain the steering, throttle, and brake commands based on the predicted waypoints. Behavior cloning [34] assumes that the expert's actions can be fully explained by observation, as it trains a model to directly map from input to output based on the training dataset (Fig. 3). However, this leads to the distribution shift problem, where the actual observations diverge from the training observations. Many latent variables impact and govern driving agent's actions in real-world scenarios. Therefore, it is essential to learn these variables effectively.

2) Direct Policy Learning: Within the context of BC, which maps sensor inputs to control commands and is limited by the training dataset, DLP aims to learn an optimal policy [33] directly that maps inputs to driving actions. The DLP algorithm obtains expert evaluations [41] during runtime to gather more training data, particularly for scenarios where the initial policy falls short. It combines an expert dataset with Imitation Learning for initial training and iteratively augments the dataset with additional trajectories collected by the trained policy. The agent can explore its surroundings and discover novel and efficient driving policies.

The online imitation learning algorithm DAGGER [79] provides robustness against cascading errors and accumulates additional training instances. Chen et al. [10] introduced automated dagger-like monitoring, where the privileged agent's supervision is collected through online learning and transformed into an agent that provides on-policy supervision. However, the main drawback of direct policy learning is the continuous need for expert access during the training process, which is both costly and inefficient.

3) Inverse Reinforcement Learning: Inverse Reinforcement Learning (IRL) [33], [74] aims to deduce the underlying specific behaviours through the reward function. Expert demonstrations $D = \{\zeta_1, \zeta_2, \zeta_3, \ldots\ldots, \zeta_n\}$ are fed into IRL. Each

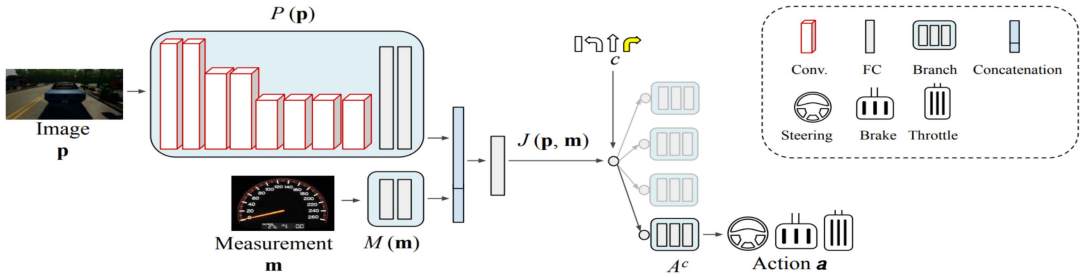


Fig. 2. Vehicle maneuvers, represented by a triplet of steering angle, throttle, and brake, depend on a high-level route navigation command (e.g., turn-left, turn-right, go-straight, continue), as well as perception data (e.g., RGB image) and vehicle state measurements (e.g., speed). These inputs guide the specific actions taken by the vehicle, enabling it to navigate the environment effectively through conditional imitation learning [47].

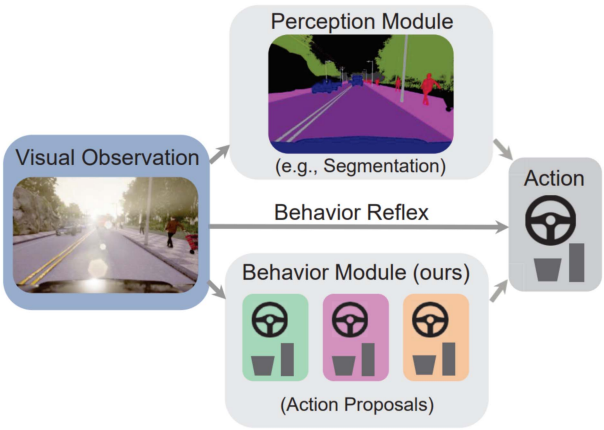


Fig. 3. Behavior cloning [34] is a perception-to-action driving model that learns behavior reflex for various driving scenarios. The agent acquires the ability to integrate expert policies in a context-dependent and task-optimized manner, allowing it to drive confidently.

$\zeta_i = \{(s_1, a_2), (s_2, a_2), \ldots..(s_n, a_n)\}$ consists of a state-action pair. The principal goal is to get the underlying reward which can be used to replicate the expert behaviour. Feature-based IRL [80] teaches the different driving styles in the highway scenario. The human-provided examples are used to learn different reward functions and capabilities of interaction with road users. Maximum Entropy (MaxEnt) inverse reinforcement learning [81] is an extension of the feature-based IRL based on the principle of maximum entropy. This paradigm robustly addresses reward ambiguity and handles sub-optimization. The major drawback is that IRL algorithms are expensive to run. They are also computationally demanding, unstable during training, and may take longer to converge on smaller datasets.

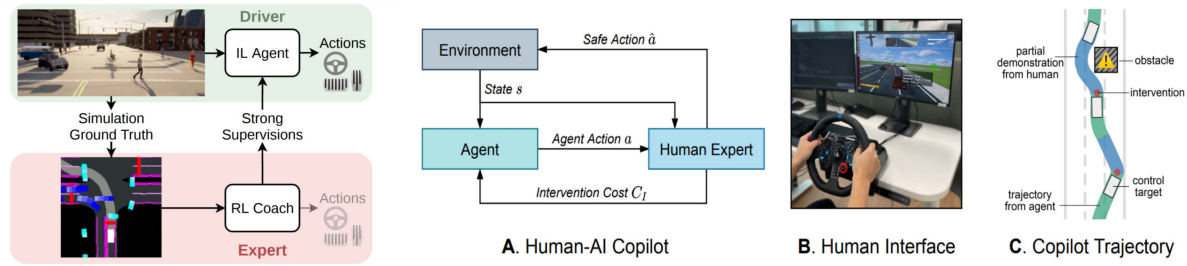
B. Reinforcement Learning

Reinforcement Learning (RL) [38], [41], [42], [72] is a promising approach to address the distribution shift problem. It aims to maximize cumulative rewards [82] over time by interacting with the environment, and the network makes driving decisions to obtain rewards or penalties based on its actions. IL cannot handle novel situations significantly different from the training dataset, RL is robust to this issue as it explores scenario under given environment. Reinforcement learning encompasses various models, including value-based models such as Deep Q-Networks (DQN) [83], actor-critic based models like Deep Deterministic Policy Gradient (DDPG) and Asynchronous Advantage Actor Critic (A3C) [83], maximum entropy models [81] such as Soft Actor Critic (SAC) [84], and policy-based optimization methods such as Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) [85].

Liang et al. [62] demonstrated the first effective RL approach for vision-based driving pipelines that outperformed the modular pipeline at the time. Their method is based on the Deep Deterministic Policy Gradient (DDPG), an extended version of the actor-critic algorithm. Chen et al. [37] uses tabular-RL to first learn an expert policy and then uses policy distillation to learn a student policy in an imitation learning approach.

Recently, Human-In-The-Loop (HITL) approaches [14], [39], [40], [41], [71] have gained attention in the literature. These approaches are based on the premise that expert demonstrations provide valuable guidance for achieving high-reward policies. Several studies have focused on incorporating human expertise into the training process of traditional RL or IL paradigms. One such example is EGPO [41], which aims to develop an expert-guided policy optimization technique where an expert policy supervises the learning agent.

HACO [40] allows the agent to explore hazardous environments while ensuring training safety. In this approach, a human expert can intervene and guide the agent to avoid potentially harmful situations or irrelevant actions (see Fig. 4(b)). Another reinforcement learning expert, Roach [39], translates bird's-eye view images into continuous low-level actions (see Fig. 4(a)). Experts can provide high-level supervision for imitation learning or reinforcement learning in general. Policies can be initially taught using imitation learning and then refined using reinforcement learning, which helps reduce the extensive training period required for RL. Jia et al. [19] utilize features extracted from Roach to learn the ground-truth action/trajectory, providing supervision in their study. Therefore, reinforcement learning provides a solution to address the challenges of imitation learning by enabling agents to actively explore and learn from their environment. There are also associated challenges, such as sample inefficiency, exploration difficulties leading to suboptimal behaviors, and difficulties generalizing learned policies to new scenarios.



(a) Roach Expert Supervision (b) Human Copilot Learning

Fig. 4. RL-based learning method for training the agent to drive optimally: (a) Illustrating the reinforcement learning expert [39] that maps the BEV to the low-level driving actions; the expert can also provide supervision to the imitation learning agent. (b) Human-in-the-loop learning [40] allows the agent to explore the environment, and in danger scenarios, the human expert takes over the control and provides the safe demonstration.

VI. Learning Domain Adaptation From Simulator to Real

Large-scale virtual scenarios can be constructed in virtual engines, enabling the collection of a significant quantity of data more readily. However, there still exist significant domain disparities between virtual and real-world data, which pose challenges in creating and implementing virtual datasets. By leveraging the principle of domain adaptation, we can extract critical features directly from the simulator and transfer the knowledge learned from the source domain to the target domain, consisting of accurate real-world data.

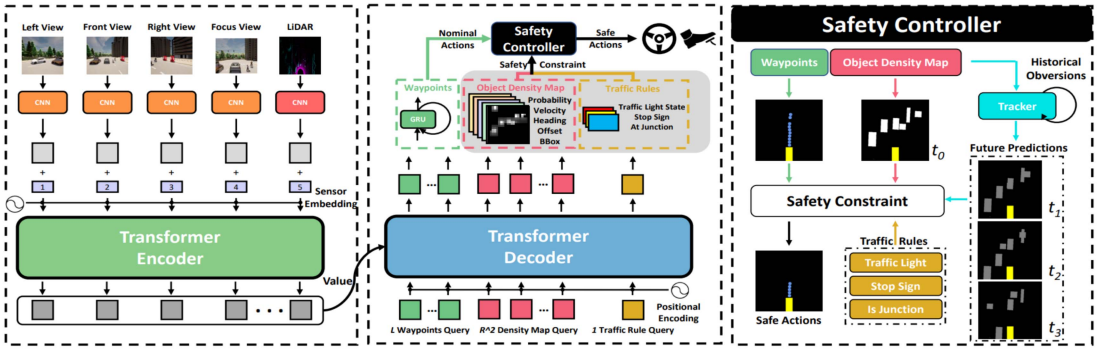
The H-Divergence framework [86] resolves the domain gap at both the visual and instance levels by adversarially learning a domain classifier and a detector simultaneously. Zhang et al. [87] propose a simulator-real interaction strategy that leverages disparities between the source domain and the target domain. The authors create two components to align differences at the global and local levels and ensure overall consistency between them. The realistic-looking synthetic images may subsequently be used to train an End-to-End model. A number of techniques rely on an open pipeline that introduces obstacles in the current environment. PlaceNet [88] places objects into the image space for detection and segmentation operations. GeoSim [89] is a geometry-aware approach that dynamically inserts objects using LiDAR and HD maps. DummyNet [90] is a pedestrian augmentation framework based on a GAN architecture that takes the background image as input and inserts pedestrians with consistent alignment.

Some works take advantage of virtual LiDAR data [91], [92], [93]. Sallab et al. [91] perform learning on virtual LiDAR point clouds from CARLA [94] and utilize CycleGAN to transfer styles from the virtual domain to the real KITTI [95] dataset. Fang et al. [92] propose a LiDAR simulator that augments real point clouds with artificial obstacles by blending them appropriately into the surroundings. Regarding planning and decision disparity, Pan et al. [96] propose learning driving policies in a simulated setting with realistic frames before applying them in the real world. Osinski et al. [97] propose a driving policy using a simulator, where a segmentation network is developed using annotated real-world data, while the driving controller is learned using synthetic images and their semantics. Mitchell et al. [98] and Stocco et al. [99] enable robust online policy learning adaptation through a mixed-reality arrangement, which includes an actual vehicle and other virtual cars and obstacles, allowing the real car to learn from simulated collisions and test scenarios.

VII. Safety

Ensuring safety in End-to-End autonomous driving systems is a complex challenge. While these systems offer high-performance potential, several considerations and approaches are essential for maintaining safety throughout the pipeline. First, training the system with diverse and high-quality data that covers a wide range of scenarios, including rare and critical situations. Hanselmann et al. [6], Chen et al. [10], Chitta et al. [7], Xiao et al. [14], and Ohn-Bar et al. [34] demonstrate that training on critical scenarios helps the system learn robust and safe behaviors and prepares it for environmental conditions and potential hazards. These scenarios include unprotected turnings at intersections, pedestrians emerging from occluded regions, aggressive lane-changing, and other safety heuristics, as shown in Fig. 5(b) and (c). Hanselmann et al. [6] focus on improving robustness by inducing adversarial scenarios (collision scenarios) and collecting an observation-waypoint dataset using experts, which is then used to fine-tune the policy.

Integrating safety constraints and rules into the End-to-End system is another vital aspect. The system can prioritize safe behavior by incorporating safety considerations during learning or post-processing system outputs. Safety constraints include a safety cost function [50], [65], [69], [74], avoiding unsafe maneuvers [11], [19], and collision avoidance strategies [13], [35], [43]. Zeng et al. [75] define the cost volume responsible for safe planning; Kendall et al. [44] propose a practical safety reward function for safety-sensitive outputs. Li et al. [40] and Hu et al. [20] demonstrate safety by utilizing the Safety Cost function that penalizes jerk, significant acceleration, and safety violations. To avoid unsafe maneuvers, Zhang et al. [36] eliminate unsafe waypoints, and Shao et al. [8] introduce InterFuser (Fig. 5(a)), which constrains only the actions within the safety



(a) InterFuser

(b) KING Lane Merger Scenario

(c) KING Collision Avoidance

Fig. 5. Demonstration of safe driving methods: (a) InterFuser [8] processes multisensorial information to detect adversarial events, which are then used by the controller to constrain driving actions within safe sets. (b) KING [6] improves collision avoidance using scenario generation. The image shows the ego vehicle (shown in red) maintaining a safe distance during a lane merge in the presence of an adversarial agent (shown in blue). (c) In the same context, the image illustrates the vehicle slowing down to avoid collision.

TABLE II
END-TO-END DRIVING TESTING TO ENSURE SAFETY

Methods	Summary	Literature
Search-based testing	Generating neuron coverage to identify false actions	[100]
	Designing an diverse and critical unsafe test cases	[6]
	Objective function to search safety sensitive output	[44], [75]
Optimization-based attack	Place the original object with an adversarial one	[88]
	Virtual obstacles to generate adversarial attack in natural environment	[22]
GAN-based attack	Generate the adversarial realistic-looking representations based on images	[101]
	Generate pedestrian augmentation from inserting pedestrians in image	[102]
	Designing an objective function to search for the diverse unsafe test cases	[8]

TABLE III
END-TO-END TESTING ORACLE MEASURES CORRECT CONTROL DECISION AT DIFFERENT SCENARIOS

Test Oracle	Detail	Literature
Metamorphic testing	The control signal should not get alter in different condition	[6]
Differential testing	The End-to-End system must give the same safe control for same scenario	[35]
Model-based oracle	Predicting the critical scenario that cause system failure	[23]

set and steers only the safest action. The above constraints ensure that the system operates within predefined safety boundaries.

Implementing additional safety modules and testing mechanisms (Tables II and III) enhances the system's safety. Real-time monitoring of the system's behavior allows for detecting abnormalities or deviations from safe operation. Hu et al. [9], Renz et al. [18], Wu et al. [13], and Hawke et al. [24] implement a planner that identifies collision-free routes, reduces possible infractions, and compensates for potential failures or inaccuracies. Renz et al. [18] use a rule-based expert algorithm for their planner, while Wu et al. [13] propose a trajectory + control model that predicts a safe trajectory over the long horizon. Hu et al. [9] also employ a goal planner to ensure safety. Codevilla et al. [23] demonstrate the system's ability to respond appropriately and return the vehicle to a safe state when encountering potential inconsistencies. Similarly, Zhao et al. [35] incorporate stop intentions to help avoid hazardous traffic situations and respond appropriately. These mechanisms ensure that the system can detect and respond to abnormal or unexpected situations, thereby reducing the risk of accidents or unsafe behavior.

Adversarial attack [22] methods, as shown in the Table II, are utilized in driving testing to evaluate the correctness of the output control signal. These testing methodologies aim to identify vulnerabilities and assess the robustness against adversaries. The End-to-End testing oracle (Table III) determines the correct control decision within a given scenario. Metamorphic testing tackles the oracle problem by verifying the consistency of the steering angle [6] across various weather and lighting conditions. It provides a reliable way to ensure that the steering

TABLE IV
POPULAR SAFETY METRICS USED FOR SAFETY EVALUATION OF DRIVING SYSTEM

Classification	Critical Metrics	Literature	Description
Temporal metrics	Time to Collision (TTC)	[103]	It defines the minimum time interval that the two agents will collide
	Worst Time to Collision (WTTC)	[104]	The WTTC metric is an extension of the traditional TTC that takes numerous traces of actors into consideration
	Time to Maneuver (TTM)	[103]	The TTM yields the latest time in the range [0, TTC] at which an expert actor may conduct a careful movement that avoids a collision
	Time to React (TTR)	[100]	TTR metric provides an approximation of the latest time before a reaction is necessary
	Time Headway (THW)	[103]	The THW measure determines the amount of time it will take an actor to get to the location of other vehicle
Non-Temporal metrics	Deceleration to Safety Time (DST)	[105]	It calculates the deceleration required to maintain the safe distance
	Stopping Distance (SD)	[106]	Minimum stopping distance at the time of deceleration
	Crash Potential Index (CPI)	[107]	It measures the probability that the vehicle cannot avoid the collision by deceleration
	Conflict Index (CI)	[107]	It estimates the collision probability and the severity factors

angle remains stable and unaffected by these factors. Differential testing [35] exposes inconsistencies among different DNN models by comparing their inference results for the same scenario. If the models produce different outcomes, it indicates unexpected behavior and potential issues in the system. The model-based oracle employs a trained probabilistic model to assess and predict potential risks [23] in real scenarios. By monitoring the environment, it can identify situations that the system may not adequately handle.

Safety metrics provide quantitative measures to evaluate the performance of autonomous driving systems and assess how well the system functions in terms of safety. Time to Collision (TTC), Conflict Index (CI), Crash Potential Index (CPI), Time to React (TTR), and others are some of the metrics that can provide additional objective comparisons between the safety performance of various approaches and identify areas that require improvement. A description of these metrics is provided in Table IV.

VIII. EXPLAINABILITY

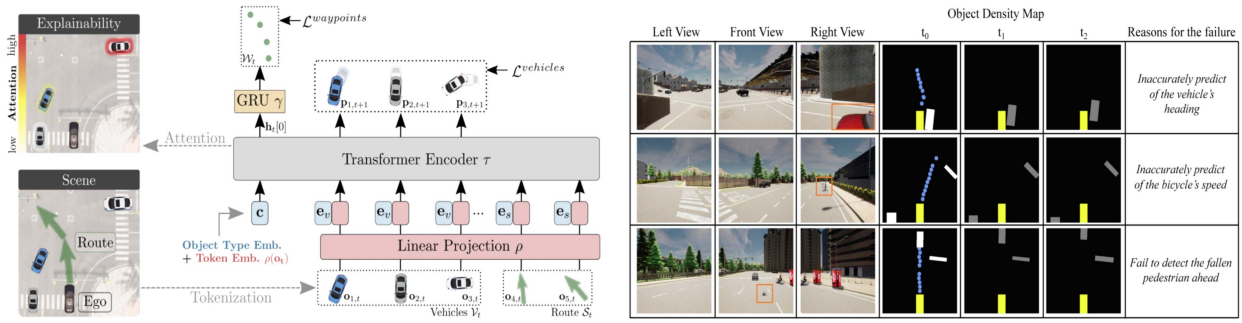
Explainability [108] refers to the ability to understand the logic of an agent and is focused on how a user interprets the relationships between the input and output of a model. It encompasses two main concepts: interpretability, which relates to the understandability of explanations, and completeness, which pertains to exhaustively defining the behavior of the model through explanations. Choi et al. [109] distinguish three types of confidence in autonomous vehicles: transparency, which refers to the person's ability to foresee and comprehend vehicle operation; technical competence, which relates to understanding vehicle performance; and situation management, which involves the notion that the user can regain vehicle control at any time. According to Haspiel et al. [110], explanations play a crucial role when humans are involved, as the ability to explain an autonomous vehicle's actions significantly impacts consumer trust, which is essential for the widespread acceptance of this technology.

In the context of explainability for end-to-end autonomous driving systems, we can categorize explanation approaches into two main types: local explanations and global explanations. A local explanation aims to describe the rationale behind the predictions of the model, represented as $y = f(x)$, when given a specific input x. On the other hand, global explanations aim to comprehensively comprehend the model's behavior by describing the underlying knowledge. As of now, there is no available research on global explanations in the context of end-to-end autonomous driving [111]. Therefore, future research should focus on addressing this gap. In the domain of local explanations, two primary approaches have gained prominence: post-hoc saliency and counterfactual explanations. Post-hoc saliency involves identifying the visual regions that have the most significant impact on the model's decisions, while counterfactual explanations help identify the contributing factors that drive specific predictions made by the model.

A. Post-Hoc Saliency Methods

A post-hoc saliency technique attempts to explain which portions of the input space have the most effect on the model's output. These approaches provide a saliency map that illustrates the locations where the model made the most significant decisions.

Post-hoc saliency methods primarily focus on the perception component of the driving architecture. Bojarski et al. [112] introduced the first post-hoc saliency approach for visualizing the impact of inputs in autonomous driving. Renz et al. [18] proposed the PlanT method (Fig. 6(a)), which utilizes an attention mechanism for post-hoc saliency visualization to provide object-level representations using the attention weights of the transformer to identify the most relevant objects. Mori et al. [113] proposed an attention mechanism that utilizes the model's predictions of steering angle and throttle. These local predictions are employed as visual attention maps and combined with learned parameters using a linear combination to make the final decision. While attention-based methods are often believed to improve the transparency of neural networks, it should be noted that learned attention weights may exhibit weak correlations with several features. The attention weights can provide accurate predictions when measuring different input features during driving. Overall, evaluating the post-hoc effectiveness of attention



(a) PlanT agent's attention (b) InterFuser failure explanation

Fig. 6. Explainability methods: (a) PlanT [18] visualization showing the attention given to the agent in various scenarios. (b) Using InterFuser [8], failure cases can be visualized by integrating three RGB views and a predicted object density map. The orange boxes indicate objects that pose a collision risk to the ego-vehicle. The object density map offers predictions for the current traffic scene (t_0) and future traffic scenes at 1-second (t_1) and 2-second (t_2) intervals.

mechanisms is challenging and often relies on subjective human evaluation.

B. Counterfactual Explanation

Saliency approaches focus on answering the 'where' question, identifying influential input locations for the model's decision. In contrast, counterfactual explanations address the 'what' question by seeking small changes in the input that alter the model's prediction. Counterfactual analysis aims to identify features X within the input x that led to the outcome $y = f(x)$ by creating a new input instance x' where X is modified, resulting in a different outcome y'. The modified input instance x' serves as the counterfactual example, and y' represents the contrasting class, such as 'What changes in the traffic scenario would cause the vehicle to stop moving?' It could be a red light.

Since the input space consists of semantic dimensions and is modifiable, assessing the causality of input components is straightforward. Li et al. [105] proposed a causal inference technique for identifying risky objects. Steex [114] developed a counterfactual method that modifies the style of the region to explain the visual model. The semantic input provides a high-level object representation, making it more interpretable compared to pixel-level representations. Bansal et al. [115] explore the underlying causes of particular outcomes by examining the ChauffeurNet model using manually crafted inputs that involve omitting particular objects.

In End-to-End driving, the steering, throttle, and brake driving outputs can be complemented with auxiliary outputs such as the occupancy and interpretable semantics to demonstrate a specific degree of counterfactual understandability. Chitta et al. [7] introduce an auxiliary output (semantics map) that employs the A* planner to address a counterfactual inquiry of "What is the possibility of a collision without braking". Shao et al. [8] designed a system, as shown in Fig. 6(b), which infers counterfactual reasoning for the potential failures with the assistance of an intermediate object density map. Sadat et al. [74] generate a probabilistic semantic occupancy map over space and time, capturing the positions of diverse road agents. Occupancy maps provide a counterfactual explanation as they act as an intermediary representation to the motion planning system, higher occupancy probabilities will discourages the maneuvers while lower occupancy will encourage them.

IX. Evaluation

The evaluation of the End-to-End system consists of open-loop evaluation and closed-loop evaluation. The open loop is assessed using real-world benchmark datasets such as KITTI [95] and nuScenes [118]. It compares the system's driving behavior with expert actions and measures the deviation. Measures such as MinADE, MinFDE [9], L2 error [20], and collision rate [75] are some of the evaluation metrics presented in Table I. In contrast the closed-loop evaluation directly assesses the system in controlled real-world or simulated settings by allowing it to drive independently and learn safe driving maneuvers.

In the open-loop evaluation of End-to-End driving systems, the system's inputs, such as camera images or LiDAR data, are provided to the system. The resulting outputs, such as steering commands and vehicle speed, are evaluated against predefined driving behaviors. The evaluation metrics commonly used in the open-loop evaluation include measures of the system's ability to follow the desired trajectory or driving behaviors, such as the mean squared error [35], L2 [9], [69] between the predicted and actual trajectories or the percentage of time the system remains within a certain distance of the desired trajectory [13]. Other evaluation metrics may also be employed to assess the system's performance in specific driving scenarios [6], [14], such as the system's capability to navigate intersections, handle obstacles, or perform lane changes. The open loop provides faster initial assessment based on functionalities and is also helpful for testing specific components or behaviors in isolation. However, they inherit drawbacks from the benchmark datasets as they cannot generalize to wider geographical distribution.

Most of the recent End-to-End systems are evaluated in closed-loop settings such as LEADERBOARD and NOCRASH [94]. Table V compares all the state-of-the-art methods on the CARLA public leaderboard. The CARLA leaderboard analyzes autonomous driving systems in unanticipated environments. Vehicles are tasked with completing a set of specified routes, incorporating risky scenarios such as unexpectedly crossing pedestrians or sudden lane changes. The leaderboard

TABLE V
CARLA AUTONOMOUS DRIVING LEADERBOARD 1.0 SUBMISSION UNTIL AUGUST 2023

Rank	Submission	DS	RC	IP	CP	CV	CL	RLI	SSI	OI	RD	AB	Type
		%	%	[0,1]	infractions/km								E/M
1	ReasonNet [16]	79.95	89.89	0.89	0.02	0.13	0.01	0.08	0.00	0.04	0.00	0.33	E
1	InterFuser [8]	76.18	88.23	0.84	0.04	0.37	0.14	0.22	0.00	0.13	0.00	0.43	E
2	TCP [13]	75.14	85.63	0.87	0.00	0.32	0.00	0.09	0.00	0.04	0.00	0.54	E
3	TF++ [56]	66.32	78.57	0.84	0.00	0.50	0.00	0.01	0.00	0.12	0.00	0.71	E
3	LAV [10]	61.85	94.46	0.64	0.04	0.70	0.02	0.17	0.00	0.25	0.09	0.10	E
4	TransFuser [7]	61.18	86.69	0.71	0.04	0.81	0.01	0.05	0.00	0.23	0.00	0.43	E
5	Latent TransFuser [7]	45.20	66.31	0.72	0.02	1.11	0.02	0.05	0.00	0.16	0.00	1.82	E
6	GRIAD [71]	36.79	61.85	0.60	0.00	2.77	0.41	0.48	0.00	1.39	1.11	0.84	E
7	TransFuser+ [7]	34.58	69.84	0.56	0.04	0.70	0.03	0.75	0.00	0.18	0.00	2.41	E
8	World on Rails [37]	31.37	57.65	0.56	0.61	1.35	1.02	0.79	0.00	0.96	1.69	0.47	E
9	MaRLn [38]	24.98	46.97	0.52	0.00	2.33	2.47	0.55	0.00	1.82	1.44	0.94	E
10	NEAT [12]	21.83	41.71	0.65	0.04	0.74	0.62	0.70	0.00	2.68	0.00	5.22	E
11	AIM-MT [12]	19.38	67.02	0.39	0.18	1.53	0.12	1.55	0.00	0.35	0.00	2.11	E
12	TransFuser [14]	16.93	51.82	0.42	0.91	1.09	0.19	1.26	0.00	0.57	0.00	1.96	E
13	CNN-Planner [116]	15.40	50.05	0.41	0.08	4.67	0.42	0.35	0.00	2.78	0.12	4.63	M
14	Learning by Cheating [33]	8.94	17.54	0.73	0.00	0.40	1.16	0.71	0.00	1.52	0.03	4.69	E
15	MaRLn [38]	5.56	24.72	0.36	0.77	3.25	13.23	0.85	0.00	10.73	2.97	11.41	E
16	CILRS [12]	5.37	14.40	0.55	2.69	1.48	2.35	1.62	0.00	4.55	4.14	4.28	E
17	CaRINA [117]	4.56	23.80	0.41	0.01	7.56	51.52	20.64	0.00	14.32	0.00	10055.99	M

Route Completion (RC), Infraction Score/penalty (IS), Driving score (DS), Collisions pedestrians (CP)/(PC), Collisions vehicles (CV), Collisions layout (CL)/(LC), Red light infractions (RLI), Red light violation (RV), Stop sign infractions (SSI), Off-road infractions (OI), Route deviations (RD), Agent blocked (AB), End-to-End Architecture (E), Modular Architecture (M).

measures how far the vehicle has successfully traveled on the given Town route within a time constraint and how many times it has incurred infractions. Several metrics provide a comprehensive understanding of the driving system, which are mentioned below:

- *Route Completion (RC):* [7], [8], [10], [13], [18] measures the percentage of the distance that an agent can complete.
- *Infraction Score/penalty (IS):* [12], [14], [36] is a geometric series that tracks infractions and aggregates the infraction penalties. It measures how often an agent drives without causing infractions.
- *Driving score (DS):* [34], [37], [39] is a primary metric calculated as the multiplication of the route completion and the infraction penalty. It measures the route completion rate weighted by infractions per route.

There are specific metrics that evaluate infractions; each metric has penalty coefficients applied every time an infraction takes place. Collisions with pedestrians, collisions with other vehicles, collisions with static elements, collisions layout, red light infractions, stop sign infractions, and off-road infractions are some of the metrics used [116]. Closed-loop evaluation provides dynamic testing adaptability where one can provide customized configuration and sensor settings. The feedback loop in it allows for iterative refinement, enabling the system to learn and improve from mistakes and experiences. However, several challenges are associated with closed-loop. These include the complexity of the initial setup and the domain gap, which might require additional fine-tuning.

X. DATASETS AND SIMULATOR

A. Datasets

In End-to-End models, the quality and richness of data are critical aspects of model training. Instead of using different hyperparameters, the training data is the most crucial factor influencing the model's performance. The amount of information fed into the model determines the kind of outcomes it produces. We summarized the self-driving dataset based on their sensor modalities, including camera, LiDAR, GNSS, and dynamics. The content of the datasets includes urban driving, traffic, and different road conditions. Weather conditions also influence the model's performance. Some datasets, such as ApolloScape [119], capture all weather conditions from sunny to snowy. The details are provided in Table VI.

B. Simulators and Toolsets

Standard testing of End-to-End driving and learning pipelines requires advanced software simulators to process information and make conclusions for their various functionalities. Experimenting with such driving systems is expensive, and conducting tests on public roads is heavily restricted. Simulation environments assist in training specific algorithms/modules before road testing. Simulators like Carla [94] offer flexibility to simulate the environment based on experimental requirements, including weather conditions, traffic flow, road agents, etc. Simulators play a crucial role in generating safety-critical scenarios and contribute to model generalization for detecting and preventing such scenarios.

Widely used platforms for training End-to-End driving pipelines are compared in Table VII. MATLAB/Simulink [150] is used for various settings; it contains efficient plot functions and has the ability to co-simulate with other software, such as CarSim [151], which simplifies the creation of different settings. PreScan [152] can mimic real-world environments, including weather conditions, which MATLAB and CarSim lack. It also supports the MATLAB Simulink interface, making modeling more effective. Gazebo [154] is well-known for its high versatility and easy connection with ROS. In contrast to the CARLA and LGSVL [153] simulators, creating a simulated environment with Gazebo requires mechanical effort. CARLA and LGSVL offer high-quality simulation frameworks that require a GPU processing unit to operate at a decent speed and frame rate. CARLA is built on the Unreal Engine, while LGSVL is based

TABLE VI
CUMULATIVE LIST OF DATASETS WITH THEIR DYNAMICS FOR END-TO-END TRAINING

Datasets	Year	Sensors Modalities							Content			Weather			Size	Location	License
		Cameras	LiDAR	GNSS	Steering	Speed, Acceleration	Navigational command	Route planner	Obstacles	Traffic	Raods	Sunny	Rain	Snow or Fog			
Udacity [120]	2016	✓	✓	✓	✓	✓			✓	✓		✓			5h	Mountain View	MIT
Drive360 [121]	2019	✓		✓	✓	✓		✓	✓			✓	✓		55h	Switzerland	Academic
Comma.ai 2016 [122]	2016	✓		✓	✓	✓						✓			7h 15min	San Francisco	CC BY-NC-SA 3.0
Comma.ai 2019 [123]	2019	✓		✓	✓	✓						✓			30h	San Jose California	MIT
DeepDrive-BDD 100 [124]	2018	✓		✓					✓	✓		✓	✓		1100h	US	Berkley
Oxford RobotCar [2]	2019	✓	✓	✓					✓	✓		✓	✓	✓	214h	Oxford	CC BY-NC-SA 4.0
HDD [125]	2018	✓	✓	✓	✓	✓	✓		✓	✓	✓				104h	San Francisco	Academic
Brain4Cars [126]	2016	✓		✓		✓			✓	✓					1180 miles	US	Academic
Li-Vi [127], [128]	2018	✓	✓	✓	✓	✓									10h	China	Academic
DDD17 [129]	2017	✓		✓	✓	✓			✓	✓		✓	✓		12h	Switzerland, Germany	CC-BY-NC-SA-4.0
A2D2 [130]	2020	✓	✓	✓	✓	✓			✓	✓		✓			390k frames	South of Germany	CC BY-ND 4.0
nuScenes [118]	2019	✓	✓	✓					✓			✓	✓		5.5h	Boston, Singapore	Non-commercial
Waymo [131]	2019	✓	✓	✓	✓	✓			✓			✓			5.5h	California	Non-commercial
H3D [30]	2019	✓	✓	✓	✓	✓						✓			N/A	Japan	Academic
HAD [132]	2019	✓		✓	✓	✓	✓			✓					30h	San Francisco	Academic
BIT [133]	2015	✓							✓			✓			9850 frames	Beijing	Academics
UA-DETRAC [134]	2015	✓							✓			✓			140k frames	Beijing, Tianjing	CC BY-NC-SA 3.0
DFG [135]	2019	✓								✓		✓	✓		7k+8k	Slovenia	CC BY-NC-SA 4.0
Bosch [136]	2017	✓								✓		✓			8334 frames	Germany	Research Only
Tsinghua-Tencent 100k [137]	2016	✓								✓		✓			30k	China	CC-BY-NC
LISA [138]	2012	✓								✓		✓			20k	California	Research Only
STSD [139]	2011	✓								✓		✓			2503 frames	Sweden	CC BY-SA 4.0
GTSRB [140]	2013	✓								✓		✓	✓		50k	Germany	CC0 1.0
KUL [141]	2013	✓								✓		✓			16k	Flanders	CC0 1.0
Caltech [142]	2009	✓							✓			✓			10 hours	California	CC4.0
CamVid [143]	2009	✓							✓	✓	✓	✓			22 min, 14 s	Cambridge	Academic
Ford [144]	2018	✓	✓						✓			✓			66 km	Michigan	CC-BY-NC-SA 4.0
KITTI [95]	2013	✓	✓	✓					✓	✓	✓	✓			43 k	Karlsruhe	Apache License 2.0
CityScapes [145]	2016	✓		✓					✓	✓		✓			20+5 K frams	Germany, France, Scotland	Apache License 2.0
Mapillary [146]	2017	✓							✓			✓	✓	✓	25000 frames	Germany	Research Only
ApolloScape [119]	2018	✓	✓						✓	✓		✓	✓	✓	147 k frames	China	Non-commercial
VERI-Wild [147]	2019	✓							✓			✓	✓		125, 280 video hours	China	Research Only
D2 -City [148]	2019	✓							✓		✓	✓	✓		10000 video	China	Research Only
DriveSeg [149]	2020	✓							✓	✓	✓	✓			500 minutes	Massachusetts	CC BY-NC 4.0

TABLE VII
PROMINENT SIMULATORS USED FOR CREATING VIRTUAL ENVIRONMENTS FOR END-TO-END SYSTEMS

Simulators	MATLAB [154]	CarSim [155]	PreScan [156]	CARLA [98]	LGSVL [157]
Sensors support	✓	✓	✓	✓	✓
Weather condition			✓	✓	✓
Camera calibration	✓		✓	✓	
Path planning	✓	✓	✓	✓	✓
Vehicle dynamics	✓	✓	✓	✓	✓
Virtual environment		✓	✓	✓	✓
Infrastructure fabrication	✓	✓	✓	✓	✓
Scenarios simulator	✓	✓	✓	✓	✓
Ground truth	✓			✓	✓
Simulator connectivity	✓	✓	✓	✓	✓
System scalability				✓	✓
Open source				✓	✓
System stable	✓	✓	✓	✓	✓
System portable	✓	✓	✓	✓	✓
API flexibility	✓	✓		✓	✓

on the Unity game engine. The API allows users to access various capabilities in CARLA and LGSVL, from developing customizable sensors to map generation. LGSVL generally links to the driving stack through various bridges, and CARLA allows built-in bridge connections via ROS and Autoware.

XI. FUTURE RESEARCH DIRECTIONS

In this section, we will highlight the possible research directions that can drive future advancements in the domain from the perspective of learning principle, safety, and explainability.

1) Learning Robustness: Current research in End-to-End autonomous driving mainly focuses on reinforcement learning (Section V-B) and imitation learning (Section V-A) methods. RL trains agents by interacting with simulated environments, while IL learns from expert agents without extensive environmental interaction. However, challenges like distribution shift in IL and computational instability in RL highlight the need for further improvements.

2) Enhanced Safety: Ensuring the behavioral safety of vehicles and accurately predicting uncertain behaviors are key aspects in safety research as discussed in Section VII. An effective system should be capable of handling various driving situations, contributing to comfortable and reliable transportation. To facilitate the widespread adoption of End-to-End approaches, it is essential to refine safety constraints and enhance their effectiveness.

3) Advancing Model Explainability: The lack of interpretability poses a new challenge for the advancement of End-to-End driving. However, ongoing efforts (Section VIII) are being made to address this issue by designing and generating interpretable features. These efforts have shown promising improvements in both performance and explainability. However, further exploration is required in global explanation strategies, including designing novel approaches to explain model actions

leading to failures and suggesting potential solutions. Future research can also explore ways to improve feedback mechanisms, allowing users to understand the decision-making process and infuse confidence in the reliability of End-to-End driving systems.

4) Collaboration Perception Systems: Vehicles can communicate directly utilizing collaborative perception to observe surroundings beyond their line of sight and field of view. This approach addresses issues related to occlusion and limited receptive fields. Cooperative or collaborative perception enables vehicles in the same area to communicate and jointly assess the scene. Cooperative perception methods [155], [156], [157] include V2V (vehicle-to-vehicle), V2I (vehicle-to-infrastructure), and V2X (vehicle-to-everything) modes. Future works should focus on enhancing the transmission efficiency within collaboration systems while safeguarding data privacy.

5) Large Language and Vision Models: Large vision models have emerged as a prominent trend in AI. By harnessing the advancements in these models, various domains can benefit from their integration. Visual prompts have become essential aids for understanding visuals across diverse domains and enhancing model capacity for interpreting visual data. Presently, SAM-Track [158] for object tracking and VIMA [159] for robot action manipulation showcase potential, implying that these large models can optimize visual recognition systems. Moreover, we can effectively utilize large language and vision models through transfer learning, domain adaptation, and fine-tuning. We can transfer insights from a larger model to a smaller one, emphasizing the importance of compact transfer of knowledge and applying it to novel tasks while upholding performance and adaptability, especially in contexts like autonomous driving. Future efforts must focus on designing large vision models tailored explicitly to autonomous driving and prompt fine-tuning to guide tasks related to perception and control.

XII. Conclusion

Over the past few years, there has been significant interest in End-to-End autonomous driving due to the simplicity of its design compared to conventional modular autonomous driving. We develop a taxonomy based on modalities, learning, and training methodology and investigate the potential of leveraging domain adaptation approaches to optimize the training process. Furthermore, the article explores evaluation framework that encompasses both open and closed-loop assessments, enabling a comprehensive analysis of system performance. To facilitate further research and development in the domain, we compile a summarized list of advancements, publicly available datasets and simulators. The article also explores potential solutions proposed by different articles regarding safety and explainability. Despite the impressive performance of End-to-End approaches, there is a need for continued exploration and improvement in safety and interpretability to achieve broader technology acceptance.

References

[1] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2722–2730.

[2] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[3] N. Akai et al., "Autonomous driving based on accurate localization using multilayer LiDAR and dead reckoning," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst.*, 2017, pp. 1–6.

[4] J. Kong, M. Pfeiffer, G. Schildbach, and F. Borrelli, "Kinematic and dynamic vehicle models for autonomous driving control design," in *Proc. IEEE Intell. Veh. Symp.*, 2015, pp. 1094–1099.

[5] P. Zhao, J. Chen, Y. Song, X. Tao, T. Xu, and T. Mei, "Design of a control system for an autonomous vehicle based on adaptive-pid," *Int. J. Adv. Robotic Syst.*, vol. 9, no. 2, p. 44, 2012.

[6] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "KING: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 335–352.

[7] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 19, 2022, doi: 10.1109/TPAMI.2022.3200245.

[8] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Proc. Conf. Robot Learn.*, 2023, pp. 726–737.

[9] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17853–17862.

[10] D. Chen and P. Krähenbühl, "Learning from all vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17222–17231.

[11] Q. Zhang, Z. Peng, and B. Zhou, "Learning to drive by watching YouTube videos: Action-conditioned contrastive policy pretraining," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 111–128.

[12] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural attention fields for end-to-end autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15793–15803.

[13] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 6119–6132.

[14] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7077–7087.

[15] P. Wu, L. Chen, H. Li, X. Jia, J. Yan, and Y. Qiao, "Policy pre-training for autonomous driving via self-supervised geometric modeling," in *Proc. Int. Conf. Learn. Representations*, 2023.

[16] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu, "ReasonNet: End-to-end driving with temporal and global reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13723–13733.

[17] Y. Xiao, F. Codevilla, D. P. Bustamante, and A. M. Lopez, "Scaling self-supervised end-to-end driving with multi-view attention learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023.

[18] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "Plant: Explainable planning transformers via object-level representations," in *Proc. CoRL Workshop Learn., Percep., Abstraction Long-Horiz. Plan.*. 2022.

[19] X. Jia et al., "Think twice before driving: Towards scalable decoders for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21983–21994.

[20] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *Proc. Eur. Conf. Comput. Vis.*2022, pp. 533–549.

[21] Q. Li, Z. Peng, H. Wu, L. Feng, and B. Zhou, "Human-ai shared control via policy dissection," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 8853–8867.

[22] H. Wu, S. Yunas, S. Rowlands, W. Ruan, and J. Wahlstrom, "Adversarial driving: Attacking end-to-end autonomous driving," in *Proc. IEEE Intell. Veh. Symp.*, 2023, pp. 1–7.

[23] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9329–9338.

[24] J. Hawke et al., "Urban driving with conditional imitation learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 251–257.

[25] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[26] L. L. Mero, D. Yi, M. Dianati, and A. Mouzakitis, "A survey on imitation learning techniques for end-to-end autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14128–14147, Sep. 2022.

[27] Z. Zhu and H. Zhao, "A survey of deep RL and IL for autonomous driving policy learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14043–14065, Sep. 2022.

[28] A. Tampuu, T. Matiisen, M. Semikin, D. Fishman, and N. Muhammad, "A survey of end-to-end driving: Architectures and training methods," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1364–1384, Apr. 2022.

[29] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," 2023, *arXiv:2306.16927*.

[30] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 9552–9557.

[31] H. Song, D. Luan, W. Ding, M. Y. Wang, and Q. Chen, "Learning to predict vehicle trajectories with model-based planning," in *Proc. Conf. Robot Learn.*, 2022, pp. 1035–1045.

[32] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 1988, pp. 305–313.

[33] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Proc. Conf. Robot Learn.*, 2020, pp. 66–75.

[34] E. Ohn-Bar, A. Prakash, A. Behl, K. Chitta, and A. Geiger, "Learning situational driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11296–11305.

[35] A. Zhao, T. He, Y. Liang, H. Huang, G. V. d. Broeck, and S. Soatto, "SAM: Squeeze-and-mimic networks for conditional visual driving policy learning," in *Proc. Conf. Robot Learn.*, 2021, pp. 156–175.

[36] J. Zhang and E. Ohn-Bar, "Learning by watching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12711–12721.

[37] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15590–15599.

[38] M. Toromanoff, E. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7153–7162.

[39] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. V. Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15222–15232.

[40] Q. Li, Z. Peng, and B. Zhou, "Efficient learning of safe driving policy via human-ai copilot optimization," in *Proc. Int. Conf. Learn. Representations*, 2021.

[41] Z. Peng, Q. Li, C. Liu, and B. Zhou, "Safe driving via expert guided policy optimization," in *Proc. Conf. Robot Learn.*, 2022, pp. 1554–1563.

[42] Y. Zhao et al., "CADRE: A cascade deep reinforcement learning framework for vision-based autonomous urban driving," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 3481–3489.

[43] J. Zhang, Z. Huang, and E. Ohn-Bar, "Coaching a teachable student," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7805–7815.

[44] A. Kendall et al., "Learning to drive in a day," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 8248–8254.

[45] A. Prakash, A. Behl, E. Ohn-Bar, K. Chitta, and A. Geiger, "Exploring data aggregation in policy learning for vision-based urban autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11763–11773.

[46] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamaki, "Multi-task learning with attention for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2902–2911.

[47] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 537–547, Jan. 2022.

[48] P. Hu, A. Huang, J. Dolan, D. Held, and D. Ramanan, "Safe local motion planning with self-supervised freespace forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12732–12741.

[49] P. Cai, S. Wang, H. Wang, and M. Liu, "Carl-Lead: LiDAR-based end-to-end autonomous driving with contrastive deep reinforcement learning," 2021, *arXiv:2109.08473*. [Online]. Available: https://api.semanticscholar.org/CorpusID:237563170

[50] W. Zeng, S. Wang, R. Liao, Y. Chen, B. Yang, and R. Urtasun, "DSDNet: Deep structured self-driving network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 156–172. [Online]. Available: https://api.semanticscholar.org/CorpusID:221112477

[51] Q. Zhang, M. Tang, R. Geng, F. Chen, R. Xin, and L. Wang, "MMFN: Multi-modal-fusion-net for end-to-end driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 8638–8643.

[52] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): Part II," *IEEE Robot. Automat. Mag.*, vol. 13, no. 3, pp. 108–117, Sep. 2006.

[53] A. Shenoi et al., "JRMOT: A real-time 3D multi-object tracker and a new large-scale dataset," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10335–10342.

[54] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7345–7353.

[55] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 641–656.

[56] P. Cai, S. Wang, H. Wang, and M. Liu, "Carl-lead: Lidar-based end-to-end autonomous driving with contrastive deep reinforcement learning," vol. abs/2109.08473, 2021.

[57] Y. Zhou et al., "End-to-end multi-view fusion for 3D object detection in LiDAR point clouds," in *Proc. Conf. Robot Learn.*, 2020, pp. 923–932.

[58] I. Sobh et al., "End-to-end multi-modal sensors fusion system for urban automated driving," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 1–9.

[59] B. Zhou, P. Krähenbühl, and V. Koltun, "Does computer vision matter for action?," *Sci. Robot.*, vol. 4, no. 30, 2019, Art. no. eaaw6661.

[60] C. Hubschneider, A. Bauer, M. Weber, and J. M. Zöllner, "Adding navigation to the equation: Turning decisions for end-to-end vehicle control," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst.*, 2017, pp. 1–8.

[61] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4693–4700.

[62] X. Liang, T. Wang, L. Yang, and E. Xing, "CIRL: Controllable imitative reinforcement learning for vision-based self-driving," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 584–599.

[63] S. Wang, J. Qin, M. Li, and Y. Wang, "FlowDriveNet: An end-to-end network for learning driving policies from image optical flow and LiDAR point flow," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 1861–1867.

[64] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2950–2958.

[65] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "LookOut: Diverse multi-future prediction and planning for self-driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16107–16116.

[66] P. Wu, L. Chen, H. Li, X. Jia, J. Yan, and Y. Qiao, "Policy pre-training for autonomous driving via self-supervised geometric modeling," in *Proc. Int. Conf. Learn. Representations*, 2023.

[67] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17252–17262.

[68] A. Hu et al., "Model-based imitation learning for urban driving," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 20703–20716.

[69] S. Casas, A. Sadat, and R. Urtasun, "MP3: A unified model to map, perceive, predict and plan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14403–14412.

[70] J. Park et al., "Object-aware regularization for addressing causal confusion in imitation learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 3029–3042.

[71] R. Chekroun, M. Toromanoff, S. Hornauer, and F. Moutarde, "GRI: General reinforced imitation and its application to vision-based autonomous driving," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021.

[72] G. Wang, H. Niu, D. Zhu, J. Hu, X. Zhan, and G. Zhou, "A versatile and efficient reinforcement learning framework for autonomous driving," 2021, *arXiv:2110.11573*.

[73] I. Kim, H. Lee, J. Lee, E. Lee, and D. Kim, "Multi-task learning with future states for vision-based autonomous driving," in *Proc. Asian Conf. Comput. Vis.*, 2020.

[74] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, predict, and plan: Safe motion planning through interpretable semantic representations," in *Proc. Eur. Conf. Comput. Vis.*2020, pp. 414–430.

[75] W. Zeng et al., "End-to-end interpretable neural motion planner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8660–8669.

[76] N. Rhinehart, R. McAllister, and S. Levine, "Deep imitative models for flexible inference, planning, and control," in *Proc. Int. Conf. Learn. Representations*, 2019.
[77] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1907–1915.
[78] M. Bain and C. Sammut, "A framework for behavioural cloning," *Mach. Intell.*, vol. 15, pp. 103–129, 1995.
[79] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist. Workshop Conf. Proc.*, 2011, pp. 627–635.
[80] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for autonomous cars that leverage effects on human actions," *Robot.: Sci. Syst.*, vol. 2, pp. 1–9, 2016.
[81] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. AAAI 23rd Nat. Conf. Artif. Intell.-Volume*, 2008, pp. 1433–1438.
[82] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[83] L. Chen, X. Hu, B. Tang, and Y. Cheng, "Conditional DQN-based motion planning with fuzzy logic for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 2966–2977, Apr. 2022.
[84] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5068–5078, Jun. 2022.
[85] X. Zhang, Y. Jiang, Y. Lu, and X. Xu, "Receding-horizon reinforcement learning approach for kinodynamic motion planning of autonomous vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 556–568, Sep. 2022.
[86] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348.
[87] H. Zhang, G. Luo, Y. Tian, K. Wang, H. He, and F.-Y. Wang, "A virtual-real interaction approach to object instance segmentation in traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 863–875, Feb. 2021.
[88] L. Zhang, T. Wen, J. Min, J. Wang, D. Han, and J. Shi, "Learning object placement by inpainting for compositional data augmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 566–581, doi: 10.1007/978-3-030-58601-0_34.
[89] Y. Chen et al., "GeoSim: Realistic video simulation via geometry-aware composition for self-driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7230–7240.
[90] A. Vobecký, D. Hurych, M. Uřičář, P. Pérez, and J. Sivic, "Artificial dummies for urban dataset augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2692–2700.
[91] A. E. Sallab, I. Sobh, M. Zahran, and M. Shawky, "Unsupervised neural sensor models for synthetic LiDAR data augmentation," 2019, *arXiv:1911.10575*.
[92] J. Fang et al., "Augmented LiDAR simulator for autonomous driving," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1931–1938, Apr. 2020.
[93] S. Manivasagam et al., "LiDARsim: Realistic LiDAR simulation by leveraging the real world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 167–11 176.
[94] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.
[95] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[96] X. Pan, Y. You, Z. Wang, and C. Lu, "Virtual to real reinforcement learning for autonomous driving," 2017, *arXiv:1704.03952*.
[97] B. Osiński et al., "Simulation-based reinforcement learning for real-world autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6411–6418.
[98] R. Mitchell, J. Fletcher, J. Panerati, and A. Prorok, "Multi-vehicle mixed reality reinforcement learning for autonomous multi-lane driving," in *Proc. 19th Int. Conf. Auton. Agents MultiAgent Syst.*, 2020, pp. 1928–1930.
[99] A. Stocco, B. Pulfer, and P. Tonella, "Mind the gap! a study on the transferability of virtual versus physical-world testing of autonomous driving systems," *IEEE Trans. Softw. Eng.*, vol. 49, no. 4, pp. 1928–1940, Apr. 2023, doi: 10.1109/TSE.2022.3202311.
[100] Y. Tian, K. Pei, S. Jana, and B. Ray, "DeepTest: Automated testing of deep-neural-network-driven autonomous cars," in *Proc. 40th Int. Conf. Softw. Eng.*, 2018, pp. 303–314.
[101] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[102] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, and P. Zhou, "Pedestrian-synthesis-GAN: Generating pedestrian data in real scene and beyond," 2018, *arXiv:1804.02047*.
[103] B. Weng, S. J. Rao, E. Deosthale, S. Schnelle, and F. Barickman, "Model predictive instantaneous safety metric for evaluation of automated driving systems," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 1899–1906.
[104] W. Wachenfeld, P. Junietz, R. Wenzel, and H. Winner, "The worst-time-to-collision metric for situation identification," in *Proc. IEEE Intell. Veh. Symp.*, 2016, pp. 729–734.
[105] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? Towards driver-centric risk assessment: Risk object identification via causal inference," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10711–10718.
[106] G. Li et al., "AV-FUZZER: Finding safety violations in autonomous driving systems," in *Proc. IEEE 31st Int. Symp. Softw. Rel. Eng.*, 2020, pp. 25–36.
[107] W. K. Alhajyaseen, "The integration of conflict probability and severity for the safety assessment of intersections," *Arabian J. Sci. Eng.*, vol. 40, pp. 421–430, 2015.
[108] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Auton. Agents Multi-Agent Syst.*, vol. 33, pp. 673–705, 2019.
[109] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *Int. J. Hum.- Comput. Interaction*, vol. 31, no. 10, pp. 692–702, 2015.
[110] J. Haspiel et al., "Explanations and expectations: Trust building in automated vehicles," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2018, pp. 119–120.
[111] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proc. 40th Int. Conf. Softw. Eng.*, 2018, pp. 303–314.
[112] M. Bojarski et al., "VisualBackProp: Efficient visualization of CNNs for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4701–4708.
[113] K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Visual explanation by attention branch network for end-to-end learning-based self-driving," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1577–1582.
[114] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering counterfactual explanations with semantics," in *Proc. Eur. Conf. Comput. Vis.*2022, pp. 387–403.
[115] M. Bansal, A. Krizhevsky, and A. Ogale, "ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst," 2018, *arXiv:1812.03079*.
[116] L. Rosero, J. Silva, D. Wolf, and F. Osório, "CNN-Planner: A neural path planner based on sensor fusion in the bird's eye view representation space for mapless autonomous driving," in *Proc. Latin Amer. Robot. Symp., Braz. Symp. Robot., Workshop Robot. Educ.*, 2022, pp. 181–186.
[117] L. A. Rosero et al., "A software architecture for autonomous vehicles: Team LRM-B entry in the first CARLA autonomous driving challenge," 2020, *arXiv:2010.12598*.
[118] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.
[119] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.
[120] "Udacity: Public driving dataset," 2017. [Online]. Available: https://github.com/udacity/self-driving-car/tree/master/datasets
[121] S. Hecker, D. Dai, and L. V. Gool, "End-to-end learning of driving models with surround-view cameras and route planners," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 435–453.
[122] E. Santana and G. Hotz, "Learning a driving simulator," 2016, *arXiv:1608.01230*.
[123] H. Schafer, E. Santana, A. Haden, and R. Biasini, "A commute in data: The comma2k19 dataset," 2018, *arXiv:1812.05752*.
[124] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2636–2645.
[125] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7699–7707.

[126] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3182–3190.
[127] Y. Chen et al., "LiDAR-video driving dataset: Learning driving policies effectively," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5870–5878.
[128] Y. Chen et al., "Lidar-video driving dataset: Learning driving policies effectively," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
[129] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "DDD20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–6.
[130] J. Geyer et al., "A2D2: AEV autonomous driving dataset," 2019. [Online]. Available: https://www.audi-electronics-venture.com/aev/web/en/driving-dataset.html
[131] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2446–2454.
[132] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10591–10599.
[133] J. Sang et al., "An improved YOLOv2 for vehicle detection," *Sensors*, vol. 18, 2018, Art. no. 4272.
[134] L. Wen et al., "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understanding*, vol. 193, 2020, Art. no. 102907.
[135] D. Tabernik and D. Skočaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1427–1440, Apr. 2020.
[136] K. Behrendt and L. Novak, "A deep learning approach to traffic lights: Detection, tracking, and classification," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 1370–1377.
[137] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2110–2118.
[138] M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for looking at traffic lights: Issues, survey, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1800–1815, Jul. 2016.
[139] F. Larsson and M. Felsberg, "Using Fourier descriptors and spatial models for traffic sign recognition," in *Proc. Scand. Conf. Image Anal.*, 2011, pp. 238–249.
[140] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: A multi-class classification competition," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 1453–1460.
[141] M. Mathias, R. Timofte, R. Benenson, and L. V. Gool, "Traffic sign recognition—how far are we from the solution?," in *Proc. Int. joint Conf. Neural Netw.*, 2013, pp. 1–8.
[142] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 304–311.
[143] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, 2009.
[144] S. Agarwal, A. Vora, G. Pandey, W. Williams, H. Kourous, and J. McBride, "Ford multi-AV seasonal dataset," *Int. J. Robot. Res.*, vol. 39, no. 12, pp. 1367–1376, 2020, doi: 10.1177/0278364920961451.
[145] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[146] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4990–4999.
[147] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3235–3243.
[148] Z. Che et al., "D^2-city: A large-scale dashcam video dataset of diverse traffic scenarios," 2019, *arXiv:1904.01975*.
[149] L. Ding, Michael Glazer, J. Terwilliger, B. Reimer, and L. Fridman, "MIT driveseg (semi-auto) dataset," *IEEE Dataport*, Jun. 3, 2020, doi: 10.21227/nb3n-kk46.
[150] T. M. Inc., "Automated driving toolbox," 2022. [Online]. Available: https://in.mathworks.com/products/automated-driving.html
[151] "Carsim adas: Moving objects and sensors," 2020. Accessed: Apr. 2023. [Online]. Available: http://bit.ly/CarSimMO
[152] Y. -J. Ou, X. -L. Wang, C. -L. Huang, J. -F. Jiang, H. -Y. Wei, and K. -S. Hsu "Application and simulation of cooperative driving sense systems using prescan software," in *Proc. Int. Conf. Inf., Commun. Eng.*, Xiamen, China, 2017, pp. 173–176, doi: 10.1109/ICICE.2017.8479296.
[153] G. Rong et al., "Lgsvl simulator: A high fidelity simulator for autonomous driving," 2020, *arXiv:2005.03778*.
[154] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IEEE Cat. No.04CH37566)*, vol. 3, Sep. 2004, pp. 2149–2154.
[155] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake LiDAR in 3D detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9243–9252.
[156] H. Xiang, R. Xu, and J. Ma, "HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," 2023, *arXiv:2304.10628*.
[157] B. Wang, L. Zhang, Z. Wang, Y. Zhao, and T. Zhou, "CORE: Cooperative reconstruction for multi-agent perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023.
[158] Y. Cheng et al., "Segment and track anything," 2023, *arXiv:2305.06558*.
[159] Y. Jiang et al., "VIMA: Robot manipulation with multimodal prompts," in *Proc. 40th Int. Conf. Mach. Learn.*, (ser. Proceedings of Machine Learning Research), Jul. 23–29, 2023, vol. 202, pp. 14 975–15 022.

Pranav Singh Chib is currently working toward the Ph.D. degree with the Computer Science and Engineering Department, Indian Institute of Technology, Roorkee, Roorkee, India. He holds a Post-Graduate Computer Science and Technology Specialization from Jawaharlal Nehru University, New Delhi, India. His research interests include machine learning, computer vision, and autonomous driving. His ongoing doctoral studies focus on contributing to advancements in autonomous driving and deep learning.

Pravendra Singh received the Ph.D. degree from the Indian Institute of Technology Kanpur, Kanpur, India. He is currently an Assistant Professor with the CSE Department, Indian Institute of Technology Roorkee, Roorkee, India. He has authored or coauthored papers at internationally reputable conferences and journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCEI, IJCV, CVPR, Proc. Eur. Conf. Comput. Vis., NeurIPS, AAAI, IJCAI, IJCV, *Pattern Recognition*, *Neural Networks*, *Knowledge-Based Systems*, *Neurocomputing*, and others. His research interests include deep learning, machine learning, computer vision, and artificial intelligence.