SMS: A Safety-Enhanced Modular Stack for Autonomous Driving

Xinshuo Weng[1], Peter Karkus[1], Yulong Cao[1], Boris Ivanovic[1], Yue Wang[1],
Yuxiao Chen[1], Apoorva Sharma[1], Marco Pavone[1,2]
[1]NVIDIA Research, [2]Stanford University
{xweng,pkarkus,yulongc,bivanovic,yuewang,yuxiaoc,apoorvas,mpavone}@nvidia.com

Abstract

End-to-end learning, which encompasses processing from sensors to actions, offers a compelling approach to autonomous driving due to its simplicity and capacity to enhance performance through data. In closed-loop driving benchmarks like the Carla Challenge, a majority of successful methods employ end-to-end learning. In contrast, the autonomous vehicle (AV) industry predominantly utilizes modular stacks with modules for detection, tracking, prediction, mapping, planning, and control, as they offer interpretability and safety guarantees. In this study, we explore the factors impeding the academic community from developing successful modular stacks and investigate how far the performance in closed-loop driving benchmarks can be advanced with a modularly-designed full AV stack. Our observation is astounding: by carefully designing all modular components, our Safety-enhanced Modular Stack (SMS), with a modern 3D object detector, tracker, and traffic detector, substantially outperforms the best publicly available end-to-end stack in the Carla offline leaderboard (84% vs. 59%). Further, SMS achieves a near-perfect score (over 97%) when provided with privileged information from the simulator. We hope that the success of SMS will motivate future research in modular stack, and augment end-to-end approaches with a significantly more performant expert.

1. Introduction

The majority of recent research on closed-loop sensor-to-control autonomous driving focuses on end-to-end learning with monolithic neural networks, while innovations on modular architectures at the system level have received much less attention. For example, in the most established closed-loop driving benchmark (the Carla challenge [3]), most methods adopt an imitation learning (IL) or reinforcement learning (RL) approach to predict actions or motion plans directly from sensors [4, 5, 7, 9, 28, 31, 39]. Despite tremendous advancements in deep policy learning and the considerable attention on this particular benchmark, to date no method has been shown to solve the task to a satisfactory level. At the same time, autonomous vehicle (AV) companies typically use modular stacks, with modules designed for detection, tracking, prediction, mapping, planning, and control [1, 12, 23, 25, 26, 32, 34, 45], in part because of their interpretability, safety guarantees and verifiability. Some of these systems have already achieved strong performance at various levels of autonomy in the real world.

We hypothesize that one reason hindering the community to pursue innovation in modular design at a large scale is the lack of mature infrastructure for a state-of-the-art (SotA) modular stack. In fact, the most popular modular AV stack developed for the Carla benchmark is Pylot [13], which does not satisfy requirements to support SotA research in closed-loop driving due to its weak performance at both the modular and system level (16.7 vs. 76.2 driving score for the SotA in the online leaderboard). Although more advanced modular stacks are developed recently, such as UniAD [16], ST-P3 [15], and DSDNet [41], they have not yet demonstrated stronger performance in full-stack closed-loop driving benchmarks such as the Carla challenge [3].

To bridge the performance gap between SotA modular and end-to-end autonomous driving stacks, and to facilitate research in modular stack development, we propose Safety-enhanced Modular Stack (SMS). The architecture is shown in Fig. 1. SMS processes sensor data (LiDAR, RGB cameras, GPS, IMU), high-definition (HD) maps, and high-level routes as inputs and generates steering, throttle, and brake commands as outputs. The driving task is decoupled into perception, motion modeling, and decision-making within SMS. To keep updated with SotA 3D perception, SMS is developed to be compatible with the OpenMM3D library [27] so we can use advanced 3D perception modules such as [20, 33, 40] for closed-loop driving. For decision making, SMS uses a sampling-based planner that supports lane change behavior, and a safety-enhanced controller that prevents collisions while not producing overly-conservative behavior.

We use the Carla closed-loop benchmark to evaluate the performance of SMS. Remarkably, when we assume perfect perception, that is, using privileged information of ground truth states from the simulator, SMS achieves a *near-perfect*, 97% driving score across all scenarios in the Carla offline leaderboard, significantly surpassing the best expert policy in prior work. Without bells and whistles, by replacing perfect perception with our learned perception module, SMS outperforms the best publicly available stack InterFuser [30]

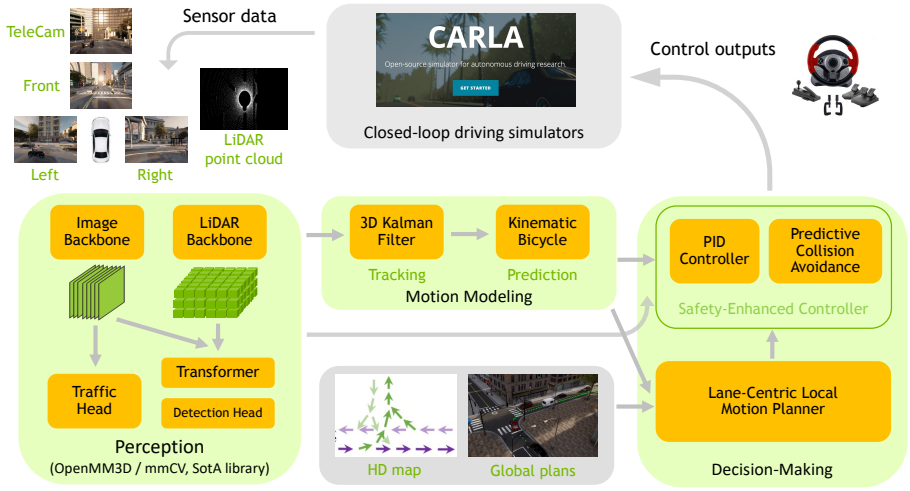

Figure 1. **SMS Architecture**. Given sensor observations, SMS outputs steering, throttle and brake signals for closed-loop driving. SMS has three main components: perception, motion modeling, and decision-making. For perception, SMS uses SotA image and LiDAR backbones available in the OpenMM3D library [27]. Features from different sensors are combined with a transformer network, and two network heads are used to detect the states of traffic signs and surrounding objects in the 3D space, respectively. The motion model predicts objects' future states using a 3D Kalman filter [38] for tracking and smoothing, and a bicycle model for motion propagation. Decision-making uses a lane-centric local motion planner that generates trajectory candidates that follow nearby lanes available in the HD map, and then chooses the best candidate based on routing and motion predictions. The local plan is converted into control signals with a PID controller. Finally, a safety layer may overwrite the control signal to ensure the resulting trajectory is collision free and complies with detected traffic signs.

by a large margin (84% vs. 59% in terms of the driving score). We hope that our results will motivate more research into modular AV stacks, offering insights into end-to-end driving benchmarks and metrics. Notably, SMS is implemented in a modular manner, making it easy to replace one module with other. We plan to release the SMS codebase to open up opportunities for system-level innovations on full-stack autonomous driving, such as joint training of multiple modules [18], and providing a significantly more performant expert for research on IL-based approaches.

2. Related Work

As opposed to the advancements of stand-alone modules for AVs such as detection and prediction driven by the Waymo [35], nuScenes [24] and KITTI [11] challenges, research at the full stack level has lagged behind. Research at the full stack level requires significantly more resources, such as (1) closed-loop sensor and behavior simulation, and accompanying benchmarks; and (2) SotA modular stack infrastructures. The Carla leaderboard [3] has established a benchmark to address the first challenge, but unfortunately, SotA modular stacks are not widely available to support stack-level research, which is the focus of this work.

Concretely, the dominant approaches in the Carla leaderboard are IL-based [4, 5, 7, 9, 28, 39], and the best available modular stack Pylot [13] does not support SotA modular design. None of the current stacks in Carla supports HD maps well, and counter-intuitively, the best map-based approach [42] has weaker performance than map-less ones on the leaderboard. In contrast to prior work, our SMS using HD maps demonstrates superior performance than all prior works in the leaderboard. In addition, we observed that the design of a safety-enhanced controller and planner that supports lane change behavior is crucial to performance, whereas we observed that prior IL-based methods do not successfully learn those behaviors in practice.

Besides planning and control, we observed that SotA perception is also not transferred to stacks for closed-loop driving, due to the lack of infrastructure. Despite the vast amount of advancements in Bird's Eye View (BEV) [20, 22] and multi-modal 3D detection [8] in the vision community, prior work in IL-based methods can only implicitly reason about the existence of objects and we empirically find a large number of collisions, let alone Pylot [13] still relies on Faster-RCNN [29] and SSD [21]. To bridge the gap between research for perception and closed-loop driving, SMS provides compatibility with OpenMM3D [27] so researchers working on perception can immediately show the impact of their work on closed-loop driving.

3. Safety-Enhanced Modular Stack

Given sensor observations from LiDAR, RGB cameras, ego vehicle's states from GPS, IMU, and also the HD map, our goal is to build a full AV stack that outputs a set of actions (brake, throttle, steer) to solve the point-to-point urban driving task. In other words, short-term goal locations are available to our stack, provided by the global route planner, and our stack aims to enable safe local motion planning and control, following the global route, and obeying traffic rules while considering the interaction with other agents. The high-level architecture of SMS is summarized in Fig. 1.

3.1. Perception

3D Object Detection. A key design decision in SMS is to leverage SotA multi-modal 3D detection for closed-loop

driving. To that end, we have designed a data interface and training loop integration with the OpenMM3D library [27]. As a result, SMS can use any of the models in OpenMM3D.

As an example, we use the standard LiDAR backbone from CenterPoint [40]/PointPillars [19], with a sparse voxel encoder and a feature pyramid network to convert point clouds into the BEV feature space. For the image stream, we follow DETR3D [33]/FUTR3D [8] to encode multi-view images into the perspective feature space. To fuse the features, a set of object queries amasses information from both the perspective and BEV feature spaces, which is then processed by a transformer decoder to generate the 3D bounding boxes.

Traffic Sign Detection. To comply with traffic rules, SMS also detects traffic lights and stop signs, and their impact on the downstream control task. The traffic sign detector shares the image backbone with 3D object detection. Subsequently, we use a few Multi-layer Perceptrons (MLPs) to classify the presence of different signs, and also their states for control implication, *e.g.*, a traffic light being red implies braking at the intersection. Notably, we include a tele-camera with a small field of view in addition to a frontal camera to detect traffic lights at a large distance.

3.2. Motion Modeling

The main goal of perception and motion modeling is not only to understand the current states of surrounding objects but also their future states, so we can plan ahead for better driving. To that end, we first aggregate the noisy frame-level detections into more stable trajectories by tracking.

3D Multi-Object Tracking. We use AB3DMOT [38] with a 3D Kalman filter [17] for tracking, due to its high computational efficiency. The output of tracking is a set of trajectories for detected objects and their associated uncertainty. Since SMS is compatible with the mm3D [27] library, it supports other SotA 3D tracking modules such as MUTR3D [43] and BEVFusion [22], although runtime is significantly slower compared to AB3DMOT and some engineering efforts such as temporal caching and model optimization are necessary to deploy them for closed-loop driving.

Motion Forecasting. As opposed to approaches developed for stand-alone trajectory prediction tasks, such as those in Waymo [35] and nuScenes [24] prediction benchmarks, SMS requires the prediction module to handle noisy input trajectories from tracking rather than ground truth past trajectories. As a result, naively applying stand-alone prediction methods might not be effective in the closed-loop setting, as also observed in prior work [36, 37]. Instead, we use simple motion models for prediction, *i.e.*, the kinematic bicycle model for vehicles, and the constant acceleration model for pedestrians. We observed that they work surprisingly well[1] in the Carla benchmark as reflected by the performance of SMS.

[1]Although SMS works well in Carla with simple motion models, we believe that extending prediction models to SotA approaches would be necessary for closed-loop driving in the real world.

3.3. Decision Making

Following the prediction of future states of surrounding agents and also the perception of the traffic signals, SMS needs to react to them safely while reaching the goal location.

Lane-Centric Motion Planning. To obtain local plans efficiently, we adopt a sampling-based planning approach. Instead of sampling a large set of trajectories covering different parts of the lane, we observed that using a lane-centric approach works surprisingly well. Specifically, SMS samples only a few plans centered at different nearby lanes. To determine which lane to take, we compute a cost based on the predicted future states of other vehicles and alignment with the global route. For example, if the current lane is blocked, and the left lane will be occupied by a fast-moving vehicle from the back, we take the right lane if it is not occupied now and in the short future according to the prediction module.

Safety-Enhanced Controller. Despite that our local planner is designed to prevent collisions, there could be inaccurate perception/prediction outputs and the emergence of new situations. Our observation is that having a safety-enhanced controller is crucial to safe driving for the Carla benchmark.

Specifically, our controller contains a basic PID component to convert the local plan to throttle, steer, and brake signals. In the meantime, SMS has a predictive collision avoidance component that takes predicted states of agents that are socially interactive to the ego vehicle and avoid collision by braking or accelerating. To prevent overly-conservative behavior, we determine if an agent is a contender to the ego based on the traffic rules. For example, when going straight into an intersection with the green light and the ego vehicle has the right of way, SMS will not brake if there is an incoming vehicle turning left (and its future trajectory collides with the ego's plans) unless the incoming vehicle is very close to the ego and still does not yield (*i.e.*, it is an adversarial agent defined in the Carla benchmark scenarios with the intention to cause difficulty to the ego vehicle).

4. Experiments

Evaluation Methodology. We use closed-loop evaluation in the Carla [10] offline leaderboard. Carla is currently the *only* closed-loop benchmark that supports full-stack urban driving, others, such as [2, 14] do not support sensor simulation or multi-agent simulation. We report the three main metrics provided by the benchmark: driving score, route completion and infraction, where the infraction summarizes all types of improper driving behaviors including collision, off-road driving, route deviation, violation of traffic rules. We use all scenario type provided by the Carla benchmark[2].

Baselines. We compare with all top methods for the Carla leaderboard as shown in Table 1 and 2, except for ReasonNet, for which the paper and code are not released yet. All

[2]Details on metrics can be found at https://leaderboard.carla.org/#task; and on scenario types at https://carlachallenge.org/challenge/nhtsa/.

Table 1. Closed-loop evaluation of the expert driver on the official 26 testing routes of the Carla offline leaderboard.

Methods	Driving score % ↑	Route completion % ↑	Infraction penalty [0, 1] ↑
LBC expert (CoRL '19) [7]	34.65	71.15	0.49
Transfuser expert (CVPR '21) [28]	45.84	85.02	0.51
Rails expert (ICCV '21) [5]	49.71	67.34	0.75
Roach expert (ICCV '21) [44]	83.00	97.00	0.85
SMS expert (**Ours**)	97.22	98.93	0.98

methods including ours are not trained on the testing routes so we can test the generalization capability of methods to new towns, new routes, and new weathers.

4.1. Expert Closed-Loop Evaluation

Collecting ground truth data is crucial to training individual modules of the stack such as perception and IL-based planners. So we first evaluate the closed-loop driving performance of expert drivers available in prior work and results are summarized in Table 1. We observed that the experts in most prior works [5, 7, 28], even with access to the ground truth perception, have a relatively weak performance, which significantly hinders the improvements of the stacks[3]. For example, we found that these expert stacks in prior work often use a simple rule-based planner which does not support lane-change behavior, and hence easily being blocked by other adversarial agents. Also, these expert stacks often reason behavior based on the current state of the world and it is not predictive in handling collisions and infractions.

In contrast to these simple experts, Roach [44] uses reinforcement learning to learn an expert and achieve higher performance by exploring in the simulator for 10 million steps. As a result, recent IL-based methods such as TCP [39] starts to use Roach to obtain the supervision for imitation and show improvements, suggesting that higher quality of expert and collected data is crucial to AV stack advancements. However, RL-based expert is not efficient to train and also lacks the verifiability and interpretability. In contrast, our SMS expert with the safety-enhanced planner and controller has demonstrated a *near-perfect* score in the Carla leaderboard, outperforming *all* available experts, and in the meantime has assured safety and interpretability due to its modular design. We hope that our expert can be beneficial to facilitate future research in closed-loop driving.

4.2. Full Stack Closed-Loop Evaluation

Besides the expert, we compared our full-stack SMS against top methods on the Carla leaderboard for closed-loop driving and the results are shown in Table 2. Compared to our expert, SMS does have a drop in driving score, from 97.22 to 84.18. We observed that this is mainly because of imperfect perception such as noisy estimation of the velocity and acceleration, which propagates to the downstream planner and controller. Overall, our SMS surpassed previous best available stack [30], improving the driving score from 58.86 to 84.18. SMS even outperforms the previous best available expert, which has access to privileged information from the simulator. Notably, the results for baselines in Table 2 are different from the online leaderboard, although we use the same metrics and scenarios for evaluation, since online leaderboard uses a set of private routes for testing.

[3]Typically, the performance of IL-based methods and also our stack is upper bounded by the expert stacks with access to the simulator.

Table 2. Closed-loop evaluation of the AV stacks on the official 26 testing routes of the CARLA offline leaderboard.

Methods	Driving score % ↑	Route completion % ↑	Infraction penalty [0, 1] ↑
LBC (CoRL '19) [7]	14.26	24.04	0.59
Transfuser (CVPR '21) [28]	24.48	65.94	0.38
World on Rails (ICCV '21) [5]	35.21	53.65	0.67
LAV (CVPR '22) [6]	43.80	77.41	0.60
TCP (NeurIPS '22) [39]	54.35	72.47	0.76
InterFuser (CoRL '22) [30]	58.86	81.33	0.69
SMS (**Ours**)	84.18	94.75	0.89

As shown in the supplementary, SMS can handle all standard NHTSA scenarios defined for driving including but not limited to performing lane change to bypass blocking objects, avoiding collision with vehicles and pedestrians, and obeying traffic rules. All the testing are conducted in new towns and new routes that are not seen during training.

5. Conclusions and Limitations

We presented a new modular stack for autonomous driving that establishes a new SotA performance in the Carla closed-loop driving benchmark, significantly outperforming previously dominant IL-based approaches. When provided with privileged information from the simulator, our stack nearly solves the challenge without infractions due to the fault of the autonomous vehicles. We hope that the success of our stack will motivate future innovation in modular stack and benefit end-to-end approaches with a more performant expert in the closed-loop driving benchmarks.

Our current approach presents two major limitations: (1) Scalability with data and joint optimization. Although we demonstrated the success of joint optimization of our perception models, we believe that joint optimization across perception, prediction and planning will yield task-aware improvements for downstream modules. (2) Demonstration of success in finer-grained level of driving and safe driving under more realistic behavior interaction. Currently, we mainly test our stack by using the default metrics provided in the Carla leaderboard, which however can only measure coarse and bottom line performance of the driving behavior, *e.g.*, having a collision or not. Also, since Carla's behavior simulation is rather simple and may not accurately represent human driving behavior in the real world, extending SMS to other more realistic closed-loop benchmarks is essential to demonstrate the generalization capability of our stack.

References

[1] Argo AI. Developing a self-driving system you

can trust, Apr. 2021. Available at https://www.argo.ai/wp-content/uploads/2021/04/ArgoSafetyReport.pdf. 1

[2] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric M. Wolff, Alex H. Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. NuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR) ADP3 Workshop*, 2021. 3

[3] Carla. Carla autonomous driving leaderboard, 2020. Available at https://leaderboard.carla.org/. 1, 2

[4] Raphael Chekroun, Marin Toromanoff, Sascha Hornauer, and Fabien Moutarde. GRI: General Reinforced Imitation and its Application to Vision-Based Autonomous Driving. *arXiv:2111.08575*, 2021. 1, 2

[5] Dian Chen, Vladlen Koltun, and Philipp Krähenbühl. Learning to drive from a world on rails. In *ICCV*, 2021. 1, 2, 4

[6] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *CVPR*, 2022. 4

[7] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning (CoRL)*, 2019. 1, 2, 4

[8] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022. 2, 3

[9] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. *CoRL*, 2017. 3

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[12] General Motors. Self-driving safety report, 2018. Available at https://www.gm.com/content/dam/company/docs/us/en/gmcom/gmsafetyreport.pdf. 1

[13] Ionel Gog, Sukrit Kalra, Peter Schafhalter, Matthew A Wright, Joseph E Gonzalez, and Ion Stoica. Pylot: A modular platform for exploring latency-accuracy tradeoffs in autonomous vehicles. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8806–8813. IEEE, 2021. 1, 2

[14] James Herman, Jonathan Francis, Siddha Ganju, Bingqing Chen, Anirudh Koul, Abhinav Gupta, Alexey Skabelkin, Ivan Zhukov, Max Kumskoy, and Eric Nyberg. Learn-to-race: A multimodal control environment for autonomous racing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9793–9802, 2021. 3

[15] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[17] R. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 1960. 3

[18] Peter Karkus, Boris Ivanovic, Shie Mannor, and Marco Pavone. Diffstack: A differentiable and modular control stack for autonomous vehicles. In *6th Annual Conference on Robot Learning*, 2022. 2

[19] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *CVPR*, 2019. 3

[20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv:2203.17270*, 2022. 1, 2

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 2

[22] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2, 3

[23] Lyft. Self-driving safety report, 2020. Available at https://2eg1kz1onwfq1djllo2xh4bb-wpengine.netdna-ssl.com/wp-content/uploads/2020/06/Safety_Report_2020.pdf. 1

[24] Motional. nuScenes Prediction Challenge, 2020. Available at https://www.nuscenes.org/prediction?externalData=all&mapData=all&modalities=Any. 2, 3

[25] Motional. Voluntary safety self-assessment, 2021. Available at https://drive.google.com/file/d/1JjfQByU_hWvSfkWzQ8PK2ZOZfVCqQGDB/view. 1

[26] NVIDIA. Self-driving safety report, 2021. Available at https://images.nvidia.com/content/self-driving-cars/safety-report/auto-print-self-driving-safety-report-2021-update.pdf. 1

[27] OpenMMLab. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1, 2, 3

[28] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[30] Hao Shao, Letian Wang, RuoBing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. *arXiv preprint arXiv:2207.14024*, 2022. 1, 4

[31] Marin Toromanoff, Mines Paristech, Valeo Dar, Emilie Wirbel, and Fabien Moutarde. End-to-End Model-Free Reinforcement Learning for Urban Driving using Implicit Affordances. *CVPR*, 2020. 1

[32] Uber Advanced Technologies Group. A principled approach to safety, 2020. Available at https://uber.app.box.com/v/UberATGSafetyReport. 1

Table 3. Closed-loop evaluation of the expert driver on the official 26 testing routes of the Carla offline leaderboard.

Methods	Driving score % ↑	Route completion % ↑	Infraction penalty [0, 1] ↑
SMS expert (**Ours**) w/o P.C.A.	77.06	91.72	0.85
SMS expert (**Ours**)	97.22	98.93	0.98

[33] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *The Conference on Robot Learning (CoRL)*, 2021. 1, 3

[34] Waymo. Safety report, 2021. Available at https://waymo.com/safety/safety-report. Retrieved on July 4, 2021. 1

[35] Waymo. Waymo Open Challenges, 2021. Available at https://waymo.com/open/. 2, 3

[36] Xinshuo Weng, Boris Ivanovic, Kris Kitani, and Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. *CVPR*, 2022. 3

[37] Xinshuo Weng, Boris Ivanovic, and Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. *IV*, 2022. 3

[38] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. *IROS*, 2020. 2, 3

[39] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *NeurIPS*, 2022. 1, 2, 4

[40] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *CVPR*, 2021. 1, 3

[41] Wenyuan Zeng, Shenlong Wang, Renjie Liao, Yun Chen, Bin Yang, and Raquel Urtasun. Dsdnet: Deep structured self-driving network. In *European Conference on Computer Vision*, pages 156–172. Springer, 2020. 1

[42] Qingwen Zhang, Mingkai Tang, Ruoyu Geng, Feiyi Chen, Ren Xin, and Lujia Wang. Mmfn: Multi-modal-fusion-net for end-to-end driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8638–8643. IEEE, 2022. 2

[43] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. *arXiv preprint arXiv:2205.00613*, 2022. 3

[44] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4

[45] Zoox. Safety report volume 2.0, 2021. Available at https://zoox.com/safety/. 1

Appendix

6. Ablation Study

Safety-Enhanced Controller. Since safety is crucial and any infraction can cause a significant reduction of scores in metrics of the Carla leaderboard and also danger for on-vehicle testing in the real world, we validate the importance of our safety-enhanced controller, specifically the predictive collision avoidance (P.C.A.) module. As shown in Table 3, if we remove the P.C.A. module and replace it with a non-predictive ones as in prior work, the driving score of our stack drops from 97.22 to 77.06, due to a number of infractions that it cannot handle without the P.C.A. module. This suggests that the design of our safety-enhanced controller is useful to closed-loop driving in the Carla leaderboard.

Table 4. Modular evaluation for traffic sign detection. Precision and recall on the 26 testing routes are reported.

SMS variants	Traffic Light Existence	Traffic Light States	Stop Sign Existence
Single-view: front camera	0.91/0.88	0.92/0.79	0.94/0.78
Single-view: left camera	0.74/0.51	0.80/0.39	0.64/0.50
Single-view: right camera	0.74/0.50	0.81/0.35	0.80/0.44
Single-view: tele-camera	0.91/0.85	0.95/0.84	0.35/0.31
Four views	0.92/0.92	0.97/0.89	0.95/0.94
Four views w. LiDAR	0.92/0.92	0.96/0.90	0.97/0.91
Joint optimization	0.92/0.94	0.96/0.94	0.96/0.95

Multi-Modal Fusion for Perception. To justify the sensor fusion for SMS, we conduct modular evaluation on the perception component. We primarily focus on the evaluation of traffic sign detection[4] because its performance has a large impact for control and closed-loop driving, *e.g.*, miss detection of a red light can lead to an infraction penalty. Results are summarized in Table 4. By comparing the first four rows (*i.e.*, single view inputs) against the four-view inputs, we observed the improved performance across the metrics. This is because there are different types of traffic signs at different places, *e.g.,* signs painted on the road and signs on the board, so information from one view might not include all places of signs. Also, adding the LiDAR sensor neither has a clear negative or positive impact on the traffic sign detection because traffic signs are mostly in the RGB images.

Joint Optimization for Perception. As a preliminary study, we jointly optimize the traffic sign detection and 3D object detection since they share the backbone and can be easily trained together. Results are shown in the last row of Table 4. We observed a minor increase in the recall without sacrificing the precision (we use the same threshold for a fair comparison). We hypothesis this might be because the joint optimization reduces the chance of overfitting of the image features to a specific task so it may increase the generalization of the image features to other towns and routes.

[4] Because our stack is integrated with the OpenMM3D library, we have tested a few modern 3D detectors such as CenterPoint and FUTR3D, and empirically find that the minor performance difference in detection does not change the closed-loop performance because jittering or miss detections often occur for objects detected at a large distance, *e.g.*, beyond 100 meters, and they are not very crucial to urban driving.