nuPlan：自律走行車のための閉ループMLベースの計画ベンチマーク

Holger Caesar Juraj Kabzan Kok Seang Tan Whye Kit Fong Eric Wolff
Alex Lang Luke Fletcher Oscar Beijbom Sammy Omari
Motional

Abstract

本研究では、世界初の自律走行のための閉ループMLベースのプランニングベンチマークを提案する。MLベースのモーションプランナーは増え続けているが、確立されたデータセットとメトリクスがないため、この分野の進歩は限定的である。既存の自律走行車運動予測ベンチマークは、長期的な計画よりも短期的な運動予測に重点を置いている。このため、これまでの研究では、L2ベースのメトリクスを用いたオープンループ評価が行われており、長期的な計画の公正な評価には適していない。我々のベンチマークは、大規模なドライビングデータセット、軽量なクローズドループシミュレータ、モーションプランニングに特化したメトリクスを導入することで、これらの制限を克服している。我々は、交通パターンが大きく異なる米国とアジアの4都市(ボストン、ピッツバーグ、ラスベガス、シンガポール)の1500時間の人間の運転データからなる高品質なデータセットを提供する。我々は、リアクティブエージェントによる閉ループシミュレーションのフレームワークを提供し、一般的な計画指標とシナリオに特化した計画指標の両方を提供する。NeurIPS 2021でデータセットを公開し、2022年初頭からベンチマーク課題を整理する予定である。

1. Introduction

大規模な人間ラベル付きデータセットとディープコンボリューショナルニューラルネットワークの組み合わせにより、ここ数年で自律走行車(AV)の知覚が目覚ましい性能向上を遂げている[9, 4]。これとは対照的に、AV計画のための既存のソリューションは、まだ主に注意深く設計されたエキスパートシステムに基づいており、新しいジオグラフィに適応するためにかなりの量のエンジニアリングを必要とし、より多くのトレーニングデータではスケールしない。適切なデータとメトリクスを提供することで、MLベースのプランニングが可能になり、完全な「ソフトウェア2.0」スタックへの道が開けると考えている。

既存の実世界のベンチマークは、プランニングではなく、予測[6, 4, 11, 8]としても知られる短期的な動き予測に焦点を当てている。これは、ハイレベルな目標の欠如、メトリクスの選択、オープンループの評価に顕著に表れている。予測は他のエージェントの行動に焦点を当て、計画は自車両の行動に関連する。

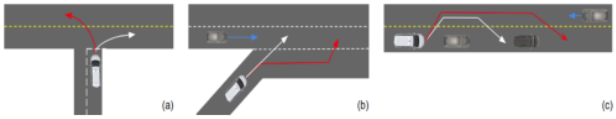

図1. 既存のベンチマークの限界を強調するために、様々な運転シナリオを示す。エゴ・ビークルの走行経路を白で、仮想的なプランナーの走行経路を赤で示す。(a) ゴールがないため、交差点で曖昧になる。(b) 変位メトリクスは、運転のマルチモーダルな性質を考慮していない。(c) オープンループ評価では、エージェントの相互作用を考慮しない。

予測は一般的にマルチモーダルであり、各エージェントについてN個の最も可能性の高い軌道を予測することを意味する。対照的に、プランニングは一般的に単一モダルであり(偶発性プランニングを除く)、我々は単一の軌道を予測する。例として、図1aでは、交差点で左折と右折が同じように起こりやすい選択肢である。予測データセットには、エージェントのハイレベルな目標を示すためのベースラインナビゲーションルートがない。図1bでは、直後または後にマージするオプションはどちらも等しく有効であるが、一般的に使用されるL2距離ベースのメトリクス(minADE、minFDE、ミス率)は、データで観察されなかったオプションにペナルティを与える。直感的には、予測された軌道と観測された軌道との距離は、マルチモーダルなシナリオでは適切な指標とはならない。図1cにおいて、追い越しを継続するか車線に戻るかの判断は、オープンループ評価では不可能な、全エージェント車両の連続動作に基づくべきである。閉ループ評価の欠如は系統的なドリフトを引き起こし、短い時間地平(3-8s)を超えて評価することを困難にする。

代わりに、これらの欠点に対処するための計画ベンチマークを提供する。我々の主な貢献は以下の通りである:

- 4都市から得られた高品質な自動ラベル付きトラックを用いた自律走行のための、既存の最大の公開実世界データセット。
- 交通ルール違反、人間の運転類似性、車両ダイナミクス、目標達成、およびシナリオベースに関連する計画指標。
- 閉ループプランナ評価プロトコルを用いた実世界データに対する最初の公開ベンチマーク。

nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles

Holger Caesar Juraj Kabzan Kok Seang Tan Whye Kit Fong Eric Wolff
Alex Lang Luke Fletcher Oscar Beijbom Sammy Omari
Motional

Abstract

In this work, we propose the world's first closed-loop ML-based planning benchmark for autonomous driving. While there is a growing body of ML-based motion planners, the lack of established datasets and metrics has limited the progress in this area. Existing benchmarks for autonomous vehicle motion prediction have focused on short-term motion forecasting, rather than long-term planning. This has led previous works to use open-loop evaluation with L2-based metrics, which are not suitable for fairly evaluating long-term planning. Our benchmark overcomes these limitations by introducing a large-scale driving dataset, lightweight closed-loop simulator, and motion-planning-specific metrics. We provide a high-quality dataset with 1500h of human driving data from 4 cities across the US and Asia with widely varying traffic patterns (Boston, Pittsburgh, Las Vegas and Singapore). We will provide a closed-loop simulation framework with reactive agents and provide a large set of both general and scenario-specific planning metrics. We plan to release the dataset at NeurIPS 2021 and organize benchmark challenges starting in early 2022.

1. Introduction

Large-scale human labeled datasets in combination with deep Convolutional Neural Networks have led to an impressive performance increase in autonomous vehicle (AV) perception over the last few years [9, 4]. In contrast, existing solutions for AV planning are still primarily based on carefully engineered expert systems, that require significant amounts of engineering to adapt to new geographies and do not scale with more training data. We believe that providing suitable data and metrics will enable ML-based planning and pave the way towards a full "Software 2.0" stack.

Existing real-world benchmarks are focused on short-term motion forecasting, also known as prediction [6, 4, 11, 8], rather than planning. This is evident in the lack of high-level goals, the choice of metrics, and the open-loop evaluation. Prediction focuses on the behavior of other agents, while planning relates to the ego vehicle behavior.

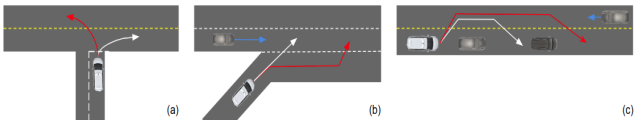

Figure 1. We show different driving scenarios to emphasize the limitations of existing benchmarks. The observed driving route of the ego vehicle in shown in white and the hypothetical planner route in red. (a) The absence of a goal leads to ambiguity at intersections. (b) Displacement metrics do not take into account the multi-modal nature of driving. (c) open-loop evaluation does not take into account agent interaction.

Prediction is typically multi-modal, which means that for each agent we predict the N most likely trajectories. In contrast, planning is typically uni-modal (except for contingency planning) and we predict a single trajectory. As an example, in Fig. 1a, turning left or right at an intersection are equally likely options. Prediction datasets lack a baseline navigation route to indicate the high-level goals of the agents. In Fig. 1b, the options of merging immediately or later are both equally valid, but the commonly used L2 distance-based metrics (minADE, minFDE, and miss rate) penalize the option that was not observed in the data. Intuitively, the distance between the predicted trajectory and the observed trajectory is not a suitable indicator in a multi-modal scenario. In Fig. 1c, the decision whether to continue to overtake or get back into the lane should be based on the consecutive actions of all agent vehicles, which is not possible in open-loop evaluation. Lack of closed-loop evaluation leads to systematic drift, making it difficult to evaluate beyond a short time horizon (3-8s).

We instead provide a planning benchmark to address these shortcomings. Our main contributions are:

- The largest existing public real-world dataset for autonomous driving with high quality autolabeled tracks from 4 cities.
- Planning metrics related to traffic rule violation, human driving similarity, vehicle dynamics, goal achievement, as well as scenario-based.
- The first public benchmark for real-world data with a closed-loop planner evaluation protocol.

Dataset	Data	Cities	Sensor Data	Type	Evaluation
Argoverse	320h	2		Pred	OL
nuPredict	5h	2	✓	Pred	OL
Lyft	1118h	1		Pred	OL
Waymo	570h	6		Pred	OL
nuPlan	1500h	4	✓	Plan	OL+CL

表1. 動き予測(Pred)とプランニング(Plan)の主要なデータセットの比較。データセットサイズ、都市数、センサーデータの可用性、データセットタイプ、オープンループ(OL)評価とクローズドループ(CL)評価のどちらを用いているかを示す。nuPredictは、nuScenes [4]データセットの予測チャレンジを指す。

2. Related Work

予測・計画データセット、シミュレーション、MLベースのプランニングに関する関連文献をレビューする。

予測データセット。表1は、我々のデータセットと関連する予測データセットとの比較を示す。Argoverse Motion Forecasting [6]は、最初の大規模予測データセットである。320時間の走行データでは、前例のないサイズであり、中心線と走行可能な領域注釈を持つ単純な意味マップを提供する。しかし、データセット中の自動ラベル付けされた軌跡は、その時の物体検出フィールドの状態や、人間がラベル付けした学習データ(113シーン)の量が不十分であるため、品質が低い。

nuScenes予測[4]チャレンジは、nuScenesデータセットから850の人間がラベル付けしたシーンから構成される。注釈は高品質でセンサーデータが提供されているが、規模が小さいため、運転バリエーションが制限される。Lyft Level 5 Prediction Dataset [11]には、6.8マイルの単一ルートからの1118hのデータが含まれている。詳細なセマンティックマップ、航空写真、動的な信号機の状態などを特徴とする。この規模は前例がないが、自動ラベル付けされたトラックはノイズが多く、地理的な多様性は限られている。Waymo Open Motion Dataset [8]は、特にエージェント間の相互作用に焦点を当てているが、オープンループ評価を用いてそうしている。データセットサイズは570hと既存のデータセットより小さいが、自動ラベル付けされたトラックは高品質である[17]。セマンティックマップと動的な信号機の状態を提供する。
これらのデータセットは、計画よりも予測に重点を置いている。本研究では、計画メトリクスと閉ループ評価を用いることで、この制限を克服することを目指す。我々は、センサーデータを提供した最初の大規模データセットである。

計画データセット CommonRoad [1]は、異なる車両モデル、コスト関数、シナリオ(目標と制約を含む)で構成される、その種類の計画ベンチマークの最初のものを提供する。事前に記録されたシナリオと対話的なシナリオの両方がある。合計5700のシナリオで、データセットの規模は、最新のディープラーニングベースの手法のトレーニングをサポートしていない。すべてのシナリオでセンサーデータがない。

シミュレーション。シミュレータは、閉ループ環境における物理、エージェント、環境条件をシミュレートする能力により、計画学習や強化学習におけるブレークスルーを可能にした。
AirSim[19]は、ドローンや自動車などのAV用の高忠実度シミュレータである。リアルタイムのハードウェア・イン・ザ・ループ・シミュレーションのために高周波数で動作する物理エンジンを搭載している。CARLA [7]は、自律型都市走行システムの訓練と検証をサポートしている。センサースイートと環境条件を柔軟に指定できる。CARLA自律走行チャレンジ[1]では、センサーデータとHDマップの異なる組み合わせを使用してウェイポイントのセットをナビゲートすることが目標である。あるいは、ユーザはシーン抽象化を使って知覚タスクを省略し、計画と制御の側面に集中することができる。この課題は概念的には我々が提案するものと似ているが、実世界のデータを使用せず、より詳細な計画指標を提供しない。

シムからリアルへの転送は、ローカライゼーション、知覚、予測、プランニング、制御など、多様なタスクのための活発な研究分野である。[21]は、合成的に学習したトラッキングモデルをKITTI [9]データセットに転送することで、シミュレーションデータと実世界データの間のドメインギャップが依然として問題であることを示している。ドメインギャップを克服するために、彼らは可視オブジェクトの実世界データとオクルージョンオブジェクトのシミュレーションデータを用いてモデルを共同で学習する。[3]は、シミュレーションから実世界のラベルなしで、ビジョンベースの車線追従走行ポリシーを実世界に転送することで、走行方法を学習する。[14]は、シミュレーションで強化学習を用いて、実世界のフルサイズ車両を制御する駆動システムを得ている。主に合成データを使用し、ラベル付けされた実世界のデータはセグメンテーションネットワークの学習にのみ現れる。

しかし、すべてのシミュレーションは系統的なバイアスをもたらすため、基本的な限界がある。実世界のセンサーをもっともらしくエミュレートするためには、例えばフォトリアリスティックなカメラ画像を生成するために、さらなる研究が必要である。

MLベースのプランニング。新しい研究分野は、実世界のデータを用いたAVのためのMLベースのプランニングである。しかし、この分野はまだ共通の入出力空間、データセット、メトリクスに収束していない。学習可能な行動と軌道プランナが共同で提案されている[18]。解釈可能なコスト関数は、知覚、予測、車両ダイナミクスのモデルの上に学習され、2つの未発表データセットでオープンループで評価される。エンドツーエンドで解釈可能なニューラルモーションプランナー[24]は、生のライダー点群と動的マップデータを入力とし、プランニングのためのコストマップを予測する。未発表のデータセットでオープンループで評価し、計画地平はわずか3秒である。ChauffeurNet[2]は、3000万もの例を使う場合でも、複雑な運転シナリオを扱うには標準的な行動クローニングでは不十分であることを発見した。彼らは、学習者が専門家の運転に摂動の形で合成されたデータにさらされることを提案し、望ましくない事象にペナルティを与え、進歩を促す追加の損失で模倣損失を補強する。

[1] See carlachallenge.org

Dataset	Data	Cities	Sensor Data	Type	Evaluation
Argoverse	320h	2		Pred	OL
nuPredict	5h	2	✓	Pred	OL
Lyft	1118h	1		Pred	OL
Waymo	570h	6		Pred	OL
nuPlan	1500h	4	✓	Plan.	OL+CL

Table 1. A comparison of leading datasets for motion prediction (Pred) and planning (Plan). We show the dataset size, number of cities, availability of sensor data, dataset type, and whether it uses open-loop (OL) or closed-loop (CL) evaluation. nuPredict refers to the prediction challenge of the nuScenes [4] dataset.

2. Related Work

We review the relevant literature for prediction and planning datasets, simulation, and ML-based planning.

Prediction datasets. Table 1 shows a comparison between our dataset and relevant prediction datasets. Argoverse Motion Forecasting [6] was the first large-scale prediction dataset. With 320h of driving data, it was unprecedented in size and provides simple semantic maps with centerlines and driveable area annotations. However, the autolabeled trajectories in the dataset are of lower quality due to the state of object detection field at the time and the insufficient amount of human-labeled training data (113 scenes).

The nuScenes prediction [4] challenge consists of 850 human-labeled scenes from the nuScenes dataset. While the annotations are high quality and sensor data is provided, the small scale limits the number of driving variations. The Lyft Level 5 Prediction Dataset [11] contains 1118h of data from a single route of 6.8 miles. It features detailed semantic maps, aerial maps, and dynamic traffic light status. While the scale is unprecedented, the autolabeled tracks are often noisy and geographic diversity is limited. The Waymo Open Motion Dataset [8] focuses specifically on the interactions between agents, but does so using open-loop evaluation. While the dataset size is smaller than existing datasets at 570h, the autolabeled tracks are of high quality [17]. They provide semantic maps and dynamic traffic light status.

These datasets focus on prediction, rather than planning. In this work we aim to overcome this limitation by using planning metrics and closed-loop evaluation. We are the first large-scale dataset to provide sensor data.

Planning datasets. CommonRoad [1] provides a first of its kind planning benchmark, that is composed of different vehicle models, cost functions and scenarios (including goals and constraints). There are both pre-recorded and interactive scenarios. With 5700 scenarios in total, the scale of the dataset does not support training modern deep learning based methods. All scenarios lack sensor data.

Simulation. Simulators have enabled breakthroughs in planning and reinforcement learning with their ability to simulate physics, agents, and environmental conditions in a closed-loop environment.

AirSim [19] is a high-fidelty simulator for AVs, such as drones and cars. It includes a physics engine that can operate at a high frequency for real-time hardware-in-the-loop simulation. CARLA [7] supports the training and validation of autonomous urban driving systems. It allows for flexible specification of sensor suites and environmental conditions. In the CARLA Autonomous Driving Challenge[1] the goal is to navigate a set of waypoints using different combinations of sensor data and HD maps. Alternatively, users can use scene abstraction to omit the perception task and focus on planning and control aspects. This challenge is conceptually similar to what we propose, but does not use real world data and provides less detailed planning metrics.

Sim-to-real transfer is an active research area for diverse tasks such as localization, perception, prediction, planning and control. [21] show that the domain gap between simulated and real-world data remains an issue, by transferring a synthetically trained tracking model to the KITTI [9] dataset. To overcome the domain gap, they jointly train their model using real-world data for visible and simulation data for occluded objects. [3] learn how to drive by transferring a vision-based lane following driving policy from simulation to the real world without any real-world labels. [14] use reinforcement learning in simulation to obtain a driving system controlling a full-size real-world vehicle. They use mostly synthetic data, with labelled real-world data appearing only in the training of the segmentation network.

However, all simulations have fundamental limits since they introduce systematic biases. More work is required to plausibly emulate real-world sensors, e.g. to generate photo-realistic camera images.

ML-based planning. A new emerging research field is ML-based planning for AVs using real-world data. However, the field has yet to converge on a common input/output space, dataset, or metrics. A jointly learnable behavior and trajectory planner is proposed in [18]. An interpretable cost function is learned on top of models for perception, prediction and vehicle dynamics, and evaluated in open-loop on two unpublished datasets. An end-to-end interpretable neural motion planner [24] takes raw lidar point clouds and dynamic map data as inputs and predicts a cost map for planning. They evaluate in open-loop on an unpublished dataset, with a planning horizon of only 3s. ChauffeurNet [2] finds that standard behavior cloning is insufficient for handling complex driving scenarios, even when using as many as 30 million examples. They propose exposing the learner to synthesized data in the form of perturbations to the expert's driving and augment the imitation loss with additional losses that penalize undesirable events and encour-

[1] See carlachallenge.org

彼らの未発表のデータセットには、60日間の連続運転に対応する2600万例が含まれている。本手法は、クローズドループとオープンループのセットアップ、および実世界で評価される。また、オープンループの評価はクローズドループと比較して誤解を招く可能性があることも示している。MP3[5]は、入力が生のライダーデータと高レベルのナビゲーションゴールである、マップレス運転へのエンドツーエンドのアプローチを提案している。未発表のデータセットでオープンループとクローズドループで評価した。マルチモーダル手法も最近の研究で研究されている[16, 20, 13]。これらのアプローチは、将来のウェイポイントや制御コマンドを予測するために、様々なモダリティ表現を融合するための様々な戦略を探求している。また、[15, 10]では、計画された軌道と観測された経路のKLダイバージェンスを用いて物体検出器を評価するニューラルプランナが用いられている。

既存の研究は、文献間で一貫性のない異なるメトリクスで評価している。TransFuser[16]は、違反の数、完了したルート距離の割合、違反乗数で重み付けされたルート完了率でその方法を評価する。違反には、他のエージェントとの衝突や赤信号の行使が含まれる。[20]はオフロード時間、オフレーン時間、衝突回数を用いてプランナーを評価し、[13, 22]は一定時間内に所定の目的地に到達する成功率を報告している。[13]はまた、ゴールまでの平均移動距離の割合を測定する別の指標を導入している。

MLベースのプランニングは非常に詳細に研究されているが、公開されているデータセットと、閉ループ評価のための共通のフレームワークを提供する標準的なメトリクスのセットがないため、この分野の進歩は制限されている。我々は、MLベースの計画データセットとメトリクスを提供することで、このギャップを埋めることを目指す。

3. Dataset

概要 Las Vegas、Boston、Pittsburgh、Singaporeから1500時間分のデータ公開を予定している。各都市は独自の推進課題を提供している。例えば、ラスベガスには、複雑なインタラクションと、一方向に最大8つの平行走行レーンを持つ多忙な交差点を持つ、賑やかなカジノのピックアップ&ドロップオフポイント(PUDO)が含まれ、ボストンのルートには、ダブルパークを好むドライバーが含まれ、ピッツバーグには、交差点の左折に関する独自のカスタム優先パターンがあり、シンガポールには左手の交通が特徴である。各都市について、セマンティックマップと効率的なマップクエリのためのAPIを提供する。データセットには、ライダー点群、カメラ画像、ローカライゼーション情報、ステアリング入力が含まれる。データセット全体で自動ラベル付けされたエージェントの軌跡を公開しているが、データセットの規模が大きい(200TB以上)ため、センサーデータのサブセットのみを利用できるようにしている。

オートラベリング。オフライン知覚システムを用いて、AVのオンライン知覚システムに課されるリアルタイムの制約なしに、大規模データセットに高精度でラベル付けを行う。PointPillars[12]とCenterPoint[23]、修正版マルチビューフュージョン(MVF++)[17]、非因果追跡を用いて、人間に近いラベリング性能を実現する。

シナリオ。シナリオベースのメトリクスを可能にするために、複雑なシナリオのタグで区間を自動的にアノテーションする。これらのシナリオには、合流、車線変更、保護または保護されていない左折または右折、自転車とのインタラクション、横断歩道またはその他の場所での歩行者とのインタラクション、近接または高加速度とのインタラクション、二重駐車車両、停止制御交差点、建設区域での走行などが含まれる。

4. Benchmarks

MLベースのプランニングの最先端技術をさらに発展させるために、以下に述べるタスクとメトリクスでベンチマークの課題を整理する。

4.1. Overview

提案手法をベンチマークデータセットに対して評価するために、ユーザはMLベースの計画コードを評価サーバに提出する。コードは提供されたテンプレートに従わなければならない。多くのベンチマークとは異なり、秘密テストセットでのクローズドループ評価を可能にするため、コードは移植性のためにコンテナ化されている。プランナーは、自動ラベル付けされた軌道、またはエンドツーエンドのオープンループアプローチでは、生のセンサーデータに対して直接動作する。特定のタイムステップについて問い合わせると、プランナーはエゴ・ビークルの計画された位置と方位を返す。提供されたコントローラは、計画された軌道を忠実に追跡しながら車両を運転する。エゴ車両の運動をシミュレートするために、あらかじめ定義された運動モデルを使用し、実際のシステムを近似する。最終的な走行軌跡は、4.2節で定義されたメトリクスに対してスコアリングされる。

4.2. Tasks

データセットに対する3つの異なるタスクを難易度を上げながら紹介する。

オープンループ。最初の課題では、人間のドライバーを模倣するように計画システムをタスク化する。タイムステップごとに、あらかじめ定義されたメトリクスに基づいて軌跡をスコアリングする。車両を制御するために使用されない。この場合、相互作用は考慮されない。

閉ループ。クローズドループの設定では、プランナは、前のケースと同様に、各タイムステップで利用可能な情報を使用して、計画された軌道を出力する。しかし、提案された軌道はコントローラの基準として使用されるため、計画システムは各タイムステップで車両の新しい状態に合わせて徐々に修正される。

age progress. Their unpublished dataset contains 26 million examples which correspond to 60 days of continuous driving. The method is evaluated in a closed-loop and an open-loop setup, as well as in the real world. They also show that open-loop evaluation can be misleading compared to closed-loop. MP3 [5] proposes an end-to-end approach to mapless driving, where the input is raw lidar data and a high-level navigation goal. They evaluate on an unpublished dataset in open and closed-loop. Multi-modal methods have also been explored in recent works [16, 20, 13]. These approaches explore different strategies for fusing various modality representations in order to predict future waypoints or control commands. Neural planners were also used in [15, 10] to evaluate an object detector using the KL divergence of the planned trajectory and the observed route.

Existing works evaluate on different metrics which are inconsistent across the literature. TransFuser [16] evaluates its method on the number of infractions, the percentage of the route distance completed, and the route completion weighted by an infraction multiplier. Infractions include collisions with other agents, and running red lights. [20] evaluates its planner using off-road time, off-lane time and number of crashes, while [13, 22] report the success rate of reaching a given destination within a fixed time window. [13] also introduces another metric which measures the average percentage of distance travelled to the goal.

While ML-based planning has been studied in great detail, the lack of published datasets and a standard set of metrics that provide a common framework for closed-loop evaluation has limited the progress in this area. We aim to fill this gap by providing an ML-based planning dataset and metrics.

3. Dataset

Overview. We plan to release 1500 hours of data from Las Vegas, Boston, Pittsburgh, and Singapore. Each city provides its unique driving challenges. For example, Las Vegas includes bustling casino pick-up and drop-off points (PUDOs) with complex interactions and busy intersections with up to 8 parallel driving lanes per direction, Boston routes include drivers who love to double park, Pittsburgh has its own custom precedence pattern for left turns at intersections, and Singapore features left hand traffic. For each city we provide semantic maps and an API for efficient map queries. The dataset includes lidar point clouds, camera images, localization information and steering inputs. While we release autolabeled agent trajectories on the entire dataset, we make only a subset of the sensor data available due to the vast scale of the dataset (200+ TB).

Autolabeling. We use an offline perception system to label the large-scale dataset at high accuracy, without the real-time constraints imposed on the online perception system of an AV. We use PointPillars [12] with CenterPoint [23], a modified version multi-view fusion (MVF++) [17], and non-causal tracking to achieve near-human labeling performance.

Scenarios. To enable scenario-based metrics, we automatically annotate intervals with tags for complex scenarios. These scenarios include merges, lane changes, protected or unprotected left or right turns, interaction with cyclists, interaction with pedestrians at crosswalks or elsewhere, interactions with close proximity or high acceleration, double parked vehicles, stop controlled intersections and driving in construction zones.

4. Benchmarks

To further the state of the art in ML-based planning, we organize benchmark challenges with the tasks and metrics described below.

4.1. Overview

To evaluate a proposed method against the benchmark dataset, users submit ML-based planning code to our evaluation server. The code must follow a provided template. Contrary to most benchmarks, the code is containerized for portability in order to enable closed-loop evaluation on a secret test set. The planner operates either on the autolabeled trajectories or, for end-to-end open-loop approaches, directly on the raw sensor data. When queried for a particular timestep, the planner returns the *planned* position and heading of the ego vehicle. A provided controller will then drive a vehicle while closely tracking the planned trajectory. We use a predefined motion model to simulate the ego vehicle motion in order to approximate a real system. The final driven trajectory is then scored against the metrics defined in Sec 4.2.

4.2. Tasks

We present the three different tasks for our dataset with increasing difficulty.

Open-loop. In the first challenge, we task the planning system to mimic a human driver. For every timestep, the trajectory is scored based on predefined metrics. It is not used to control the vehicle. In this case, no interactions are considered.

Closed-loop. In the closed-loop setup the planner outputs a *planned* trajectory using the information available at each timestep, similar to the previous case. However, the proposed trajectory is used as a reference for a controller, and thus, the planning system is gradually corrected at each timestep with the new state of the vehicle. While the new state of the vehicle may not coincide with that of

車両の新しい状態が記録された状態と一致しない可能性があり、カメラビューやライダー点群が異なるが、センサーデータのワーピングや新しいビュー合成は行わない。このセットでは、2つのタスクを区別する。非反応性閉ループ課題では、他のエージェントの動作を仮定せず、観測されたエージェントの軌跡を単純に使用する。11]で示されているように、閉ループシミュレーションにおける介入の大部分は、例えば、車両が自車両と素朴に衝突するなどの非反応性によるものである。反応的閉ループタスクでは、エゴ・ビークルのように追跡される他の全てのエージェントの計画モデルを提供する。

4.3. Metrics

我々は、メトリクスを、すべてのシナリオに対して計算される共通メトリクスと、事前に定義されたシナリオに合わせたシナリオベースのメトリクスの2つのカテゴリに分割する。

Common metrics.

- トラフィックルール違反は、一般的なトラフィックルールへの準拠を測定するために使用される。他のエージェントとの衝突率、オフロード軌道の割合、リードエージェントまでの時間差、衝突までの時間、エージェントを通過する間の相対速度を、通過距離の関数として計算する。

- 人間の運転類似度は、人間と比較して操縦満足度を定量化するために使用される(例:縦方向の速度誤差、縦方向の停止位置誤差、横方向の位置誤差)。さらに、得られたジャーク/加速度を人間レベルのジャーク/加速度と比較する。

- 車両ダイナミクスは、ライダーの快適性と軌道の実現可能性を定量化する。ライダーの快適性は、ジャーク、加速度、操舵速度、車両振動によって測定される。実現可能性は、同じ基準の定義済みの限界に違反することによって測定される。

- 目標達成度は、L2距離を用いて、地図上の目標ウェイポイントに向かうルートの進捗を測定する。

シナリオベースのメトリクス。第3節のシナリオタグに基づき、難易度の高い操縦のための追加メトリクスを使用する。車線変更については、衝突までの時間、目標車線上のリード/リアエージェントまでの時間ギャップを測定し、スコア化する。歩行者と自転車の相互作用については、歩行者の位置を区別しながら、通過相対速度を定量化する。さらに、横断歩道と無防備な曲がり角(通行権)について、プランナーと人間による意思決定の一致度を比較する。

コミュニティからのフィードバック。ここに示した指標は最初の提案であり、網羅的なリストを形成していないことに注意。

我々はコミュニティと緊密に協力し、コミュニティ全体のコンセンサスを得るために、新しいシナリオとメトリクスを追加する。同様に、主要なチャレンジメトリックについては、メトリックの加重和、あらかじめ定義された閾値を超えるメトリック違反の加重和、メトリックの階層など、複数のオプションが見られる。我々は、この分野を前進させるメトリクスを定義するために、コミュニティが我々と協力することを呼びかける。

5. Conclusion

本研究では、AVのための最初のMLベースの計画ベンチマークを提案した。既存の予測ベンチマークとは異なり、我々はゴールベースの計画、計画メトリクス、閉ループ評価に焦点を当てる。共通のベンチマークを提供することで、自律走行における最終的なフロンティアの一つであるMLベースのプランニングの進歩への道を開くことを期待している。

References

[1] Matthias Althoff, Markus Koschi, and Stefanie Manzinger. CommonRoad: Composable benchmarks for motion planning on roads. In *Proc. of the IEEE Intelligent Vehicles Symposium*, 2017. 2

[2] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *RSS*, 2019. 2

[3] Alex Bewley, Jessica Rigley, Yuxuan Liu, Jeffrey Hawke, Richard Shen, Vinh-Dieu Lam, and Alex Kendall. Learning to drive from simulation without real world labels. In *ICRA*, 2019. 2

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2

[5] Sergio Casas, Abbas Sadat, and Raquel Urtasun. MP3: A unified model to map, perceive, predict and plan. In *CVPR*, 2021. 3

[6] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 1, 2

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. *CoRR*, 2017. 2

[8] Scott Ettinger, Shuyang Cheng, and Benjamin Caine et al. Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. *arXiv preprint arXiv:2104.10133*, 2021. 1, 2

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11):1231–1237, 2013. 1, 2

[10] Yiluan Guo, Holger Caesar, Oscar Beijbom, Jonah Philion, and Sanja Fidler. The efficacy of neural planning metrics: A

the recorded state, leading to different camera views or lidar point clouds, we will not perform any sensor data warping or novel view synthesis. In this set, we distinguish between two tasks. In the *Non-reactive closed-loop* task we do not make any assumptions on other agents behavior and simply use the observed agent trajectories. As shown in [11], the vast majority of interventions in closed-loop simulation is due to the non-reactive nature, e.g. vehicles naively colliding with the ego vehicle. In the *reactive closed-loop* task we provide a planning model for all other agents that are tracked like the ego vehicle.

4.3. Metrics

We split the metrics into two categories, common metrics, which are computed for every scenario and scenario-based metrics, which are tailored to predefined scenarios.

Common metrics.

- *Traffic rule violation* is used to measure compliance with common traffic rules. We compute the rate of collisions with other agents, rate of off-road trajectories, the time gap to lead agents, time to collision and the relative velocity while passing an agents as a function of the passing distance.

- *Human driving similarity* is used to quantify a maneuver satisfaction in comparison to a human, e.g. longitudinal velocity error, longitudinal stop position error and lateral position error. In addition, the resulting jerk/acceleration is compared to the human-level jerk/acceleration.

- *Vehicle dynamics* quantify rider comfort and feasibility of a trajectory. Rider comfort is measured by jerk, acceleration, steering rate and vehicle oscillation. Feasibility is measured by violation of predefined limits of the same criteria.

- *Goal achievement* measures the route progress towards a goal waypoint on the map using L2 distance.

Scenario-based metrics. Based on the scenario tags from Sec. 3, we use additional metrics for challenging maneuvers. For *lane change*, time to collision and time gap to lead/rear agent on the target lane is measured and scored. For *pedestrian/cyclist interaction*, we quantify the passing relative velocity while differentiating their location. Furthermore, we compare the *agreement between decisions made by a planner and human* for crosswalks and unprotected turns (right of way).

Community feedback. Note that the metrics shown here are an initial proposal and do not form an exhaustive list. We will work closely with the community to add novel scenarios and metrics to achieve consensus across the community. Likewise, for the main challenge metric we see multiple options, such as a weighted sum of metrics, a weighted sum of metric violations above a predefined threshold or a hierarchy of metrics. We invite the community to collaborate with us to define the metrics that will drive this field forward.

5. Conclusion

In this work we proposed the first ML-based planning benchmark for AVs. Contrary to existing forecasting benchmarks, we focus on goal-based planning, planning metrics and closed-loop evaluation. We hope that by providing a common benchmark, we will pave a path towards progress in ML-based planning, which is one of the final frontiers in autonomous driving.

References

[1] Matthias Althoff, Markus Koschi, and Stefanie Manzinger. CommonRoad: Composable benchmarks for motion planning on roads. In *Proc. of the IEEE Intelligent Vehicles Symposium*, 2017. 2

[2] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *RSS*, 2019. 2

[3] Alex Bewley, Jessica Rigley, Yuxuan Liu, Jeffrey Hawke, Richard Shen, Vinh-Dieu Lam, and Alex Kendall. Learning to drive from simulation without real world labels. In *ICRA*, 2019. 2

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2

[5] Sergio Casas, Abbas Sadat, and Raquel Urtasun. MP3: A unified model to map, perceive, predict and plan. In *CVPR*, 2021. 3

[6] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 1, 2

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. *CoRR*, 2017. 2

[8] Scott Ettinger, Shuyang Cheng, and Benjamin Caine et al. Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. *arXiv preprint arXiv:2104.10133*, 2021. 1, 2

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11):1231–1237, 2013. 1, 2

[10] Yiluan Guo, Holger Caesar, Oscar Beijbom, Jonah Philion, and Sanja Fidler. The efficacy of neural planning metrics: A

meta-analysis of PKL on nuscenes. In *IROS Workshop on Benchmarking Progress in Autonomous Driving*, 2020. 3

[11] John Houston, Guido Zuidhof, and Luca Bergamini et al. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020. 1, 2, 4

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

[13] Eraqi Hesham M., Mohamed N. Moustafa, and Jens Honer. Conditional imitation learning driving considering camera and lidar fusion. In *NeurIPS*, 2020. 3

[14] Blazej Osinski, Adam Jakubowski, Pawel Ziecina, Piotr Milos, Christopher Galias, Silviu Homoceanu, and Henryk Michalewski. Simulation-based reinforcement learning for real-world autonomous driving. In *ICRA*, 2020. 2

[15] Jonah Philion, Amlan Kar, and Sanja Fidler. Learning to evaluate perception models using planner-centric metrics. In *CVPR*, 2020. 3

[16] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[17] Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. *arXiv preprint arXiv:2103.05073*, 2021. 2, 3

[18] Abbas Sadat, Mengye Ren, Andrei Pokrovsky, Yen-Chen Lin, Ersin Yumer, and Raquel Urtasun. Jointly learnable behavior and trajectory planning for self-driving vehicles. In *IROS*, 2019. 2

[19] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017. 2

[20] Ibrahim Sobh, Loay Amin, Sherif Abdelkarim, Khaled Elmadawy, Mahmoud Saeed, Omar Abdeltawab, Mostafa Gamal, and Ahmad El Sallab. End-to-end multi-modal sensors fusion system for urban automated driving. In *NeurIPS*, 2018. 3

[21] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. *arXiv preprint arXiv:2103.14258*, 2021. 2

[22] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. Multimodal end-to-end autonomous driving. *arXiv preprint arXiv:1906.03199*, 2019. 3

[23] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *arXiv preprint arXiv:2006.11275*, 2020. 3

[24] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *CVPR*, 2021. 2

meta-analysis of PKL on nuscenes. In *IROS Workshop on Benchmarking Progress in Autonomous Driving*, 2020. 3

[11] John Houston, Guido Zuidhof, and Luca Bergamini et al. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020. 1, 2, 4

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

[13] Eraqi Hesham M., Mohamed N. Moustafa, and Jens Honer. Conditional imitation learning driving considering camera and lidar fusion. In *NeurIPS*, 2020. 3

[14] Blazej Osinski, Adam Jakubowski, Pawel Ziecina, Piotr Milos, Christopher Galias, Silviu Homoceanu, and Henryk Michalewski. Simulation-based reinforcement learning for real-world autonomous driving. In *ICRA*, 2020. 2

[15] Jonah Philion, Amlan Kar, and Sanja Fidler. Learning to evaluate perception models using planner-centric metrics. In *CVPR*, 2020. 3

[16] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[17] Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. *arXiv preprint arXiv:2103.05073*, 2021. 2, 3

[18] Abbas Sadat, Mengye Ren, Andrei Pokrovsky, Yen-Chen Lin, Ersin Yumer, and Raquel Urtasun. Jointly learnable behavior and trajectory planning for self-driving vehicles. In *IROS*, 2019. 2

[19] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017. 2

[20] Ibrahim Sobh, Loay Amin, Sherif Abdelkarim, Khaled Elmadawy, Mahmoud Saeed, Omar Abdeltawab, Mostafa Gamal, and Ahmad El Sallab. End-to-end multi-modal sensors fusion system for urban automated driving. In *NeurIPS*, 2018. 3

[21] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. *arXiv preprint arXiv:2103.14258*, 2021. 2

[22] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. Multimodal end-to-end autonomous driving. *arXiv preprint arXiv:1906.03199*, 2019. 3

[23] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *arXiv preprint arXiv:2006.11275*, 2020. 3

[24] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *CVPR*, 2021. 2