自律走行のための占有予測ガイド付きニューラルプランナー

Haochen Liu, Zhiyu Huang, and Chen Lv*, *Senior Member, IEEE*

概要- 複雑な交通シナリオにおける周囲の交通参加者のスケーラブルな将来状態を予測することは、安全で実現可能な意思決定を可能にするため、自律走行車にとって重要な能力である。学習ベースの予測とプランニングにおける最近の成功は、2つの主要な課題を導入している:環境のための正確な共同予測の生成と、プランニング目的のための予測ガイダンスの統合。これらの課題に対処するために、我々はOPGPと呼ばれる2段階の統合ニューラルプランニングフレームワークを提案する。OPGPは、占有予測からの共同予測ガイダンスを組み込んでいる。予備計画フェーズでは、統一されたTransformer構造の中で、共有された相互作用、シーンコンテキスト、アクターダイナミクスを考慮しながら、模倣学習目標に基づいて、様々なタイプのトラフィックアクターの予測占有率を同時に出力する。その後、変換された占有予測は、Frenet座標の下での安全で滑らかなプランニングにさらに情報を提供するために、最適化を導く。大規模な実走行データセットを用いてプランナを訓練し、オープンループ構成で検証する。提案するプランナは、強力な学習ベースの手法を凌駕し、占有予測ガイダンスにより性能が向上する。

I. Introduction

複数の交通参加者(エージェント)が意思決定モジュールに情報を提供するために、頑健で社会的に適合性のある共同先物を正確に予測することは、自律走行システム(ADS)にとって極めて重要な能力である[1]-[3]。しかし、予測とプランニングを統合することは、いくつかの要因から大きな課題がある。まず、ADSは、社会的に相互作用する異種交通参加者の組み合わせを特徴とする、膨大な数の複雑な交通シーンを処理しなければならない[4]。第二に、運動予測器は、近傍の多数の交通アクターの将来の状態の共同パターンを管理しなければならない[5]。さらに、経路レベルのプランニング決定には、ADSの安全で滑らかなプランニング性能を実行するために、共同予測から実現可能なガイダンスが必要である。

最適な予測ガイダンスを達成するために、予測研究の大部分は、選択された周囲の参加者の将来の位置の共同シーケンスを直接マッピングするマルチエージェント軌跡予測(MATP)に焦点を当てている[6]、[7]。しかし、このアプローチは探索コストによって妨げられ、探索コストは参加者数とともに線形に増加し、限界運動予測では指数関数的に増加する[8]。最近、研究者は占有グリッドの予測に目を向けている[9]。視覚表現として、占有予測は複数の参加者に対してより効率的でスケーラブルな形式を提供する。

Code is available at: https://github.com/georgeliu233/OPGP

H. Liu, Z. Huang, and C. Lv are with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, 639798, Singapore. (E-mails: {haochen002, zhiyu001}@e.ntu.edu.sg, lyuchen@ntu.edu.sg)

This work was supported in part by the A*STAR under MTC IRG Grant (No. M22K2c0079) and the SUG-NAP Grant (No. M4082268.050) of Nanyang Technological University, Singapore.

*Corresponding author: C. Lv

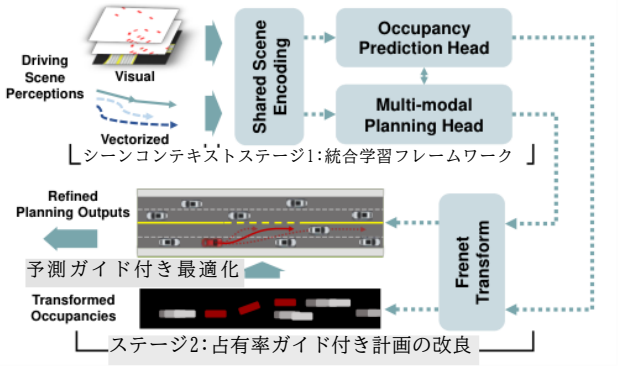

図1. 提案する2段階占有予測誘導型ニューラルプランナ(OPGP)。統合学習ベースのフレームワークから予測される占有率とマルチモーダル計画出力。次に、予測ガイド付き最適化により、変換された予測と計画から計画を洗練させる。

さらに、様々な参加者の共同将来を一目で予測することができ[10]、計画決定を統合する際にMATPと比較してオクルージョン予測可能性の安全性が向上する。しかし、共同占有率予測の文脈で、すべての対話的アクターの社会的認識を捉えることは、依然として未解決の問題である。

逆に、学習ベースのモーションプランニングは、ますます注目されている[11]。統一されたニューラルネットワークを使用することで、多様な運転シナリオに対応できる素晴らしい結果を達成した。それにもかかわらず、学習ベースの計画の頑健性と安全性は不安定性によって損なわれる可能性がある。この問題に対処するために、予測ガイダンスを統合することが、この学習構造の標準的なアプローチとして浮上している[12]-[14]。隣接する予測に基づいて能動的な応答と運動制約を学習することで、より安全で、よりロバストで、社会的に適合した計画システムが期待される[15]。一般的な方法の1つは、周囲のアクターの将来の軌跡を計画に利用することで予測を統合することである[16], [17]。関節軌道のモデル化にMATPを使用することは困難であるにもかかわらず、予測ガイダンスのソースとして占有率を探求した研究は限られている。共同将来占有率の予測は、計画のためのニューラルネットワーク内に視覚ベースの予測特徴を統合する際に、かなりの課題を提示する。さらに、情報量の多い予測占有率を実現可能な計画結果に変換することは、さらなる調査が必要な課題である。

これらの課題に取り組むために、図1に示すように、将来の占有率と運動計画のための共同予測と予測ガイダンスを統合する2段階学習ベースのフレームワーク(OPGPと命名)を提案する。第一段階では、Transformerのバックボーン上に、占有予測ガイド付きプランニングの統合ネットワークが確立される。我々の以前の研究[18]に基づき、すべてのタイプのトラフィック参加者の占有予測は同時に出力される、

Occupancy Prediction-Guided Neural Planner for Autonomous Driving

Haochen Liu, Zhiyu Huang, and Chen Lv*, *Senior Member, IEEE*

***Abstract*— Forecasting the scalable future states of surrounding traffic participants in complex traffic scenarios is a critical capability for autonomous vehicles, as it enables safe and feasible decision-making. Recent successes in learning-based prediction and planning have introduced two primary challenges: generating accurate joint predictions for the environment and integrating prediction guidance for planning purposes. To address these challenges, we propose a two-stage integrated neural planning framework, termed OPGP, that incorporates joint prediction guidance from occupancy forecasting. The preliminary planning phase simultaneously outputs the predicted occupancy for various types of traffic actors based on imitation learning objectives, taking into account shared interactions, scene context, and actor dynamics within a unified Transformer structure. Subsequently, the transformed occupancy prediction guides optimization to further inform safe and smooth planning under Frenet coordinates. We train our planner using a large-scale, real-world driving dataset and validate it in open-loop configurations. Our proposed planner outperforms strong learning-based methods, exhibiting improved performance due to occupancy prediction guidance.**

I. Introduction

Accurately predicting robust, socially compatible joint futures for multiple traffic participants (agents) to inform the decision-making module is a crucial capability for autonomous driving systems (ADS) [1]–[3]. However, integrating prediction and planning presents significant challenges due to several factors. First, ADS must process a vast array of complex traffic scenes, each featuring combinations of socially interactive heterogeneous traffic participants [4]. Second, the motion predictor must manage joint patterns of future states for numerous traffic actors in the vicinity [5]. Moreover, path-level planning decisions necessitate feasible guidance from joint predictions to execute safe and smooth planning performances for ADS.

To achieve optimal prediction guidance, the majority of prediction research focuses on multi-agent trajectory prediction (MATP), which directly maps the joint sequences of future locations for selected surrounding participants [6], [7]. However, this approach is hindered by the search cost, which grows linearly with the number of participants and exponentially for marginal motion predictions [8]. Recently, researchers have turned to forecasting occupancy grids [9].

As a visual representation, occupancy prediction provides a more efficient and scalable form for multiple participants.

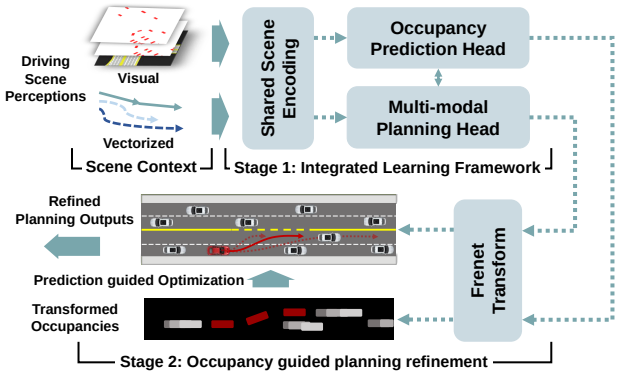

Fig. 1. Proposed two-stage occupancy prediction-guided neural planner (OPGP). Predicted occupancy and multi-modal planning outputs from an integrated learning-based framework. We then refine the planning from transformed prediction and planning via prediction-guided optimizations.

In addition, it can predict the joint future for various participants at a single glance [10], offering enhanced safety for occlusion predictability compared to MATP when integrating planning decisions. However, capturing the social awareness of all interactive actors in the context of joint occupancy forecasting remains an open question.

Conversely, learning-based motion planning has gained increasing attention [11]. Its ability to handle diverse driving scenarios has achieved impressive results through the use of a unified neural network. Nevertheless, the robustness and safety of learning-based planning can be undermined by instability. To address this issue, integrating prediction guidance has emerged as a standard approach for this learning structure [12]–[14]. By learning active responses and motion constraints based on neighboring predictions, a safer, more robust, and socially compliant planning system is anticipated [15]. One common method involves integrating predictions by using the future trajectories of surrounding actors for planning [16], [17]. Despite the difficulties of using MATP to model joint trajectories, limited research has explored occupancy as a source of prediction guidance. Predicting joint future occupancy presents considerable challenges in integrating visual-based prediction features within the neural network for planning. Furthermore, transforming informative predicted occupancy into feasible planning outcomes is an issue that warrants further investigation.

To tackle these challenges, as shown in Fig. 1, we propose a two-stage learning-based framework (named OPGP) that integrates joint predictions for future occupancy and motion planning with prediction guidance. In the first stage, an integrated network of occupancy prediction-guided planning is established upon Transformer backbones. Building upon our previous work [18], occupancy predictions for all types of traffic participants are output simultaneously,

Code is available at: https://github.com/georgeliu233/OPGP

H. Liu, Z. Huang, and C. Lv are with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, 639798, Singapore. (E-mails: {haochen002, zhiyu001}@e.ntu.edu.sg, lyuchen@ntu.edu.sg)

This work was supported in part by the A*STAR under MTC IRG Grant (No. M22K2c0079) and the SUG-NAP Grant (No. M4082268.050) of Nanyang Technological University, Singapore.

*Corresponding author: C. Lv

視覚的特徴とベクトル化されたコンテキストの両方に対する相互作用の認識を考慮する。一方、エンコードされたシーン特徴と占有率はプランナヘッドで共有され、条件付きで問い合わせられ、ADSの軌道に対してマルチモーダルな運動計画を行う。第2段階は、最適化可能な方法で、モーションプランニングの洗練のための占有予測からの明示的なガイダンスをモデル化することに焦点を当てる。より具体的には、Frenet空間[19]において、変換された占有予測を用いた計画洗練のための最適化パイプラインを構築する。我々の提案するフレームワークの貢献は以下のようにまとめられる：

1）2段階の占有予測ガイド付きニューラルプランナパイプラインを提案する。第一段階は、統合された占有予測ガイド付き運動計画フレームワークを含む。予測された占有率は、交通参加者間の社会的相互作用とシーンコンテキストを考慮した、Transformerベースの構造におけるマルチモーダルな計画軌道と組み合わされる。

2）第2段階では、計画性能をさらに向上させるために、計画洗練のための変換された占有予測ガイド付き最適化を設計する。

3）大規模な実走行データセットを用いて2段階のフレームワークを検証し、提案するパイプラインは説得力のある性能を達成する。

II. Related Work

A. 関節運動予測

関節運動予測のための学習ベースの技術が開発されつつあり、非常に効果的であることが証明されている[1]。これは、ディープニューラルネットワーク、特にTransformersとGNN[20]が、複数の相互作用する参加者と多様なシーンコンテキストを含む複雑なトラフィックシナリオを処理する能力に起因する。エージェント中心の手法は、検出された各交通参加者に固定されたマルチエージェント単位の未来軌道(MATP)を予測する。DenseTNT [6]、M2I [21]、HEAT [22]は、GNNバックボーンを持つ各アクターについて、マージナルな組み合わせで共同予測をスコアリングする。各アクターに対して精度を示す一方で、時折矛盾が生じる計算コストも高くなる。ジョイント法[23]、[24]は、ヒートマップのスコアリングやサンプリングから、すべてのアクターの将来の分布を直接推定する。しかし、これらの方法は、最大数の予測を必要とし、特に混雑した都市部では、複数のエージェントで線形的なコスト増加を持つ。将来の運動を予測するための占有グリッドの使用は、自律走行で数年前から採用されている技術である。注目すべき例として、ChauffeurNet [25]があり、これは占有マップを利用して将来の動きを予測し、行動計画を容易にする。StopNet[10]はさらに、スケーラブルでリアルタイムな予測のために、統合された軌跡と占有予測を行う。我々の以前の方法[18]に基づき、予測学習時に学習フレームワークに計画情報を組み込む。我々は、すべてのタイプの交通参加者を含むように、占有予測パイプラインを改良し、シーンコンテキスト内でのそれらの相互作用を考慮し、プランニングをガイドするための強化を行った。

B. 予測ガイダンスによる動作計画

モーションプランニングは、時代とともに広範な研究が行われてきた、確立された分野である。経路最適化[26]、サンプリング[27]、そして最近では学習ベースの技術[11]、[28]など、様々なアプローチによって発展してきた。しかし、インタラクティブで洗練された交通で安全に巡航するためには、参加者の行動の予測状態からさらにガイダンスが必要である。PiP[29]は、サンプリングに基づく計画再スケジューリングを繰り返し条件付き予測を行うが、これは限界的な未来のみを考慮し、生成された計画経路によって制限される。DIPP [16]は、微分可能な計画目標を共同軌道予測と結びつけ、応答的な計画を可能にする。これまでの手法では、エージェント中心の方法で予測ガイダンスに焦点を当てていたが、我々は、成長するアクターに対するスケーラビリティと不変性の観点から、シーン中心のガイダンスを追求する。Prediction Net[30]は、プランニングと制御のための後処理された軌跡を持つ占有予測ガイダンスを示す。しかし、交通シミュレーションのための占有率予測に重点を置いており、特に計画のための統合を設計していない。MP3[31]やInterfuser[32]のようなエンドツーエンドの手法は、生の視覚入力から予測と計画を直接マッピングする。とはいえ、彼らは知覚部分を予測に統一することに重点を置いており、計画出力は単に広範な軌跡記憶から取り出されるだけである。我々の研究では、提案するニューラルプランナの両段階を導くために、占有率予測を活用する。具体的には、予測特徴は学習ベースの計画デコーダによって対話的に照会され、その結果得られた予測は実行可能な方法で計画を洗練するために利用される。

III. Methodology

A. 問題の定式化

図1に示すように、占有予測ガイド付きニューラルプランナは2つのステージから構成される。第一段階では、占有予測ガイダンスを用いた運動計画決定は、マルチタスク学習パラダイムとして定式化することができる。シーンコンテキストSを入力として、統合学習フレームワークfは、占有予測O^1:Tと、将来の水平線Tにおけるマルチモーダル模倣計画軌道Y^e1:Tを同時に出力する。その後、第2段階は、予測と計画の両方にFrenet変換[19]を利用し、洗練された計画出力τ^*を生成し、定義されたコストCを最小化する、達成可能な予測ガイド付き最適化を達成する：

$$\begin{aligned} \hat{\mathbf{Y}}^{e}_{1:T}, \hat{\mathbf{O}}_{1:T} &= f(\mathcal{S}|\theta), \\ \tau^* &= \arg\min_{\tau} \; C(\hat{\mathbf{Y}}^{\mathbf{T}}_{1:T}, \hat{\mathbf{O}}^{\mathbf{T}}_{1:T}, \mathcal{S}), \end{aligned} \tag{1}$$

ここで、θはモデルパラメータを表す。より具体的には、入力シーンコンテキストSは、履歴地平T_hの下で複数のモダリティから構成され、詳細な定式化は以下の通りである：1）視覚的特徴：特定のシナリオの下での交通参加者の空間的・時間的状態を表す、履歴占有グリッド$O_{Th:0}$とラスタライズされたBEVロードマップM [33]から組み合わせを構築する。

taking into account interaction awareness for both visual features and vectorized context. Meanwhile, encoded scene features and occupancy are shared and conditionally queried in the planner head, which conducts multi-modal motion planning for the ADS's trajectories. The second stage focuses on modeling explicit guidance from occupancy prediction for motion planning refinement in an optimization-feasible manner. More specifically, we construct an optimization pipeline in Frenet space [19] for planning refinement using transformed occupancy predictions. The contributions of our proposed framework are summarized as follows:

1) We propose a two-stage occupancy prediction guided neural planner pipeline. The first stage involves an integrated occupancy prediction-guided motion planning framework. Predicted occupancy is combined with multi-modal planning trajectories in Transformer-based structures, taking into account the social interactions among traffic participants and scene context.
2) In the second stage, we design a transformed occupancy prediction guided optimization for planning refinement to further enhance planning performance.
3) We validate the two-stage framework on a large-scale real-world driving dataset, and the proposed pipeline achieves compelling performance.

II. Related Work

A. Joint motion prediction

A growing number of learning-based techniques have been developed for joint motion predictions, proving highly effective [1]. This can be attributed to the capability of deep neural networks, particularly Transformers and GNNs [20], to handle intricate traffic scenarios involving multiple interacting participants and diverse scene contexts. Agent-centric methods predict multi-agent-wise future trajectories (MATP) anchored on each detected traffic participant. DenseTNT [6], M2I [21], and HEAT [22] score joint predictions in combinations of marginal ones for each actor with GNN backbones. While they demonstrate accuracy for each actor, they also raise computational costs with occasional inconsistency. Joint methods [23], [24] directly estimate the future distribution of all actors from heatmap scoring or sampling. However, these methods require a maximum number of predictions and have linear cost growth with multiple agents, especially in crowded urban areas. The use of occupancy grids to forecast future motions has been a technique employed in autonomous driving for several years. One notable example is ChauffeurNet [25], which utilizes occupancy maps to predict future movements and facilitate behavior planning. StopNet [10] further conducts integrated trajectory and occupancy predictions for scalable and real-time predictions. Building on our previous method [18], we incorporate planning information into the learning framework during prediction training. We have refined our occupancy prediction pipeline to include all types of traffic participants and consider their interactions within the scene context for enhancement to guide planning.

B. Motion planning with prediction guidance

Motion planning is a well-established field that has received extensive study over time. It has developed through various approaches such as path optimization [26], sampling [27], and more recently, learning-based techniques [11], [28]. However, cruising safely in interactive and sophisticated traffic requires further guidance from the predicted states of participants' behaviors. PiP [29] makes conditional predictions iteratively on sampling-based planning rescheduling, which only considers marginal futures and is limited by generated planning paths. DIPP [16] ties differentiable planning objectives with joint trajectory predictions, enabling responsive planning. While previous methods focus on prediction guidance in an agent-centric manner, we seek scene-centric guidance in terms of scalability and invariance for growing actors. PredictionNet [30] demonstrates occupancy prediction guidance with post-processed trajectories to inform planning and control. However, it focuses on occupancy prediction for traffic simulation and does not specifically design integration for planning. End-to-end methods such as MP3 [31] and Interfuser [32] directly map prediction and planning from raw visual inputs. Nevertheless, they focus more on unifying the perception part into prediction, and the planning outputs are simply retrieved from extensive trajectory storage. In our work, we leverage occupancy prediction to guide both stages of our proposed neural planner. Specifically, the prediction features are queried interactively by the learning-based planning decoder, and the resulting predictions are utilized to refine the planning in a feasible manner.

III. Methodology

A. Problem Formulation

As shown in Fig. 1, the occupancy prediction guided neural planner consists of two stages. In the first stage, motion-planning decisions with occupancy prediction guidance can be formulated as a multi-task learning paradigm. Conditioned on scene context $\mathcal{S}$ as inputs, the integrated learning framework f simultaneously outputs occupancy predictions $\hat{\mathbf{O}}1:T$ and multi-modal imitated planning trajectories $\hat{\mathbf{Y}}^e 1:T$ in future horizons T. Subsequently, the second stage utilizes a Frenet transformation [19] for both prediction and planning to achieve attainable prediction-guided optimization that generates refined planning outputs τ^*, minimizing the defined costs C. Mathematically, the two-stage task is formulated as:

$$\begin{aligned} \hat{\mathbf{Y}}^e_{1:T}, \hat{\mathbf{O}}_{1:T} &= f(\mathcal{S}|\theta), \\ \tau^* &= \arg\min_{\tau} \; C(\hat{\mathbf{Y}}^{\mathbf{T}}_{1:T}, \hat{\mathbf{O}}^{\mathbf{T}}_{1:T}, \mathcal{S}), \end{aligned} \tag{1}$$

where θ denotes the model parameters. More specifically, the input scene context $\mathcal{S}$ comprises multiple modalities under historical horizon T_h with detailed formulations as follows: **1) Visual features:** We construct a combination from the historical occupancy grids $\mathbf{O}_{T_h:0}$ and rasterized BEV roadmap $\mathcal{M}$ [33] representing the spatial-temporal status of traffic participants under a specific scenario. Each

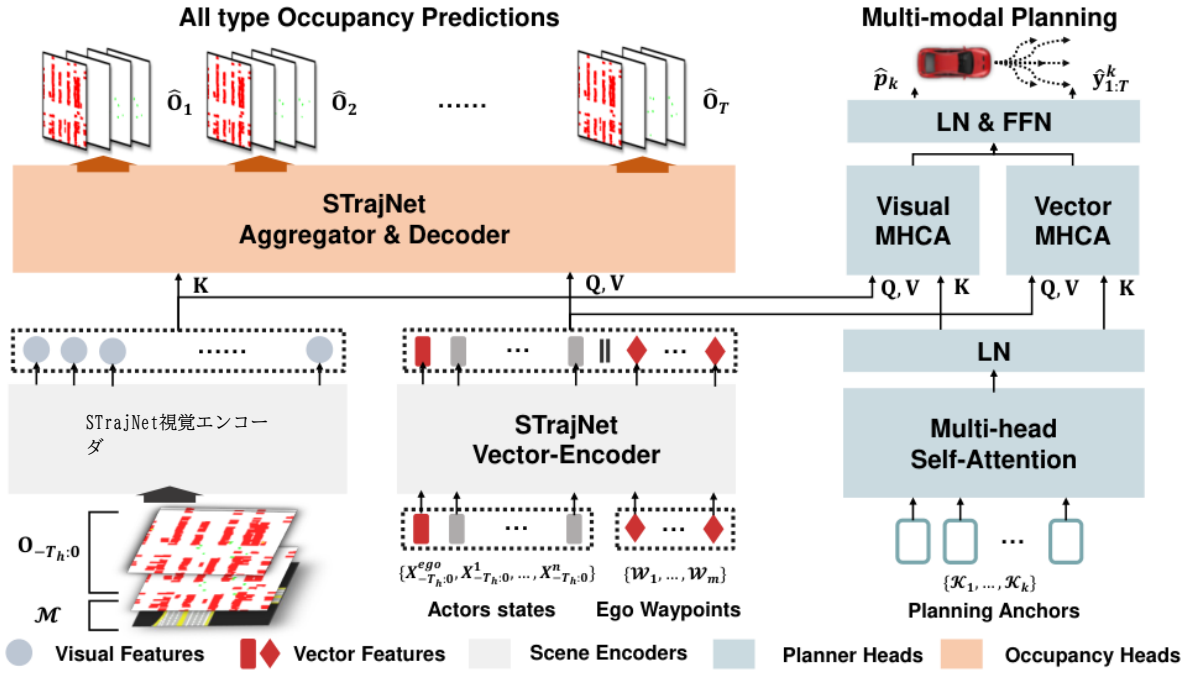

図2. OPGPにおける占有予測・計画学習統合フレームワークの概要。エンコードされたシーン特徴は、すべてのタイプの占有予測$\hat{O}_{1:T}$に対する占有予測ヘッドと、後続ステージの初期計画出力$\hat{Y}^{e}_{1:T}$に対するマルチモーダル計画ヘッドによって分配される。

各占有率O_tは3種類の交通参加者(車両、歩行者、自転車)を個別に含み、さらに現在未見のアクターに対する予測$\hat{O}_{1:T}$のためのオクルージョン次元を追加する。2) ベクトル特徴: ベクトル化されたコンテキストは、まず、自律走行車$X_{-Th:0} = X^{ego}_{-Th:0}$, $X^{1}_{-Th:0}$, – –, $X^{n}_{-Th:0}$ を中心とする最大n_x + 1人のアクターからの動的コンテキストに関係し、それぞれが過去の運動シーケンスを表す: (x, y, v_x, v_y, θ)。ワンホットエンコーディングはそれぞれのタイプに関連する。自律走行車(エゴ)の地図コンテキストを提供するために、m個のウェイポイントW_mと参照ルートIを含むLanelets[34]のグラフ検索を使用して地図データを収集する。各地図セグメントには、座標情報だけでなく、道路タイプ、制限速度、近くの信号機の現在の状態などの道路特徴も含まれる。

B. 統合予測・計画フレームワーク

図2は、我々の学習ベースのフレームワークの全体構造を示しており、占有予測とプランニングを統合し、3つのモジュールから構成されている。まず、マルチモーダルなシーン入力Sに対して、我々の先行研究の別符号化を採用する。エンコードされた視覚特徴量とベクトル特徴量は、その後、占有予測ヘッドとプランナーヘッドに配信される。占有ヘッドについては、以前の集約と復号のパイプラインを維持し、エンコードされたベクトルと視覚的特徴の後期融合を行う。予測ヘッドは最終的に、未来の地平線の各秒について、すべてのタイプの占有予測$\hat{O}_{1:T}$を出力する。マルチモーダル計画ヘッドについては、Transformerデコーダのパラダイムに従い、マルチモーダル経路計画$\hat{Y}^{e}_{1:T}$に対して単層復号ヘッドを採用する。

より具体的には、最初にk個の計画アンカーK_1, –, K_kのセットを導入し、計画ヘッドのk個のモードをガイドする。各アンカーKは、静的な終点クラスタか、計画解読を発見的に導く学習可能な埋め込みのための動的な終点クラスタのどちらかである。これらのアンカーは、自己アテンション後に符号化された視覚的特徴とベクトル特徴に対するコンテキストクエリQとして機能する。

マルチヘッドクロスアテンションモジュール(MHCA)では、エゴ車両の視覚的コンテキスト、ベクトル化されたダイナミクス、シーントポロジーから、インタラクションを考慮した特徴を適応的に得ることを目的とする。これは、連結された注意ヘッドを考慮すると、パディングされた視覚的特徴の長さが注意スコアリングを疎にするためである。その後のフィードフォワードネットワーク(FFN)の後、各モードは計画経路$\hat{y}^{k}_{1:T}$とそれに対応する模倣尤度の確率$\hat{p}_k$の出力にデコードする。

我々は、第一段階の学習フレームワークをマルチタスク学習パラダイムとして定式化し、大規模データセットを用いて目的をオフラインで共同更新する。占有予測では、各グリッドセルがある将来の地平線の下で占有確率を予測するため、アクターのタイプごとに占有サンプルのバランスをとるためにフォーカルロス[35]を採用する:

$$\mathcal{L}_{pred} = \frac{1}{HWTU} \sum_{x,y,t,u}^{H,W,T,U} \mathrm{FL}(\hat{\mathbf{O}}^{u}_{t}(x,y), \mathbf{O}^{u}_{t}(x,y)), \quad (2)$$

ここで、H, Wは占有スケール、Uはアクタータイプを表す。マルチモーダル計画では、模倣は

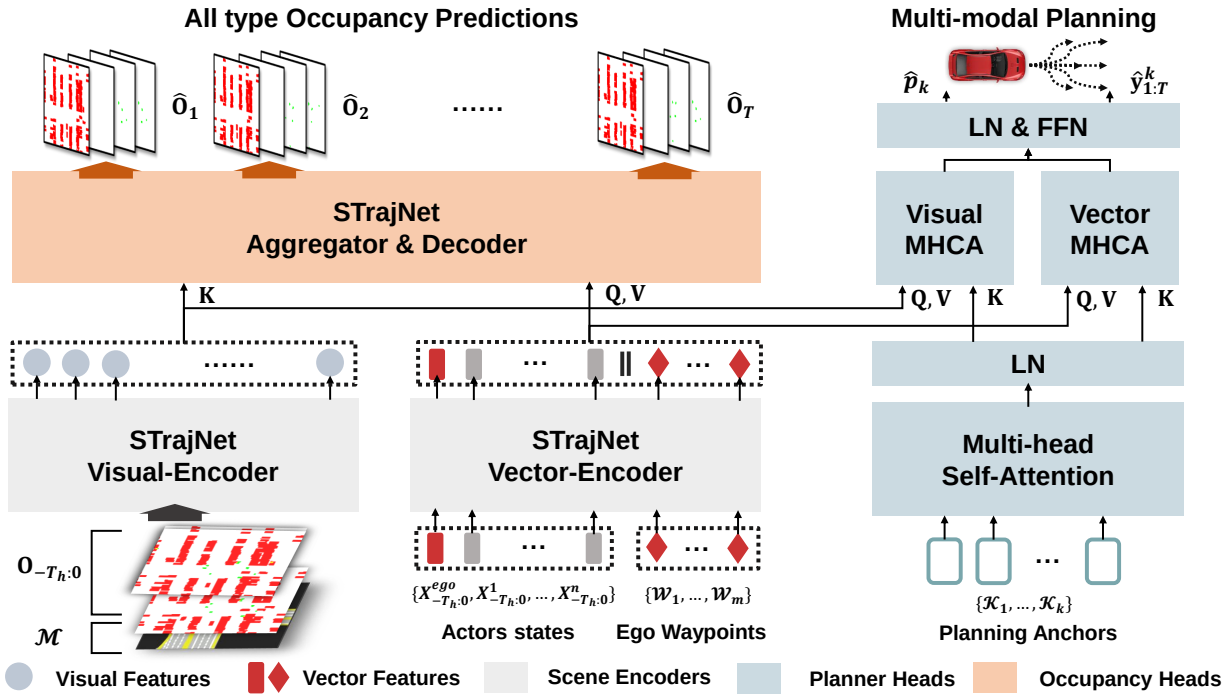

Fig. 2. An overview of the integrated occupancy prediction and planning learning framework in OPGP. Building on our previous work, visual and vector scene inputs $\mathcal{S}$ are separately encoded in scene encoders; the encoded scene features are distributed by occupancy prediction heads for all type occupancy predictions $\hat{\mathbf{O}}_{1:T}$ and multi-modal planning heads for initial planning outputs $\hat{\mathbf{Y}}^e_{1:T}$ for the subsequent stage.

occupancy $\mathbf{O}_t$ separately contains 3 types of traffic participants (vehicles, pedestrians, and cyclists), and we further add an occlusion dimension for predictions $\hat{\mathbf{O}}_{1:T}$ for currently unseen actors. **2) Vector features:** The vectorized context initially concerns the dynamic context from a maximum of n_x+1 actors centered on the autonomous vehicle $\mathbf{X}_{-T_h:0} = X^{ego}_{-T_h:0}, X^1_{-T_h:0}, \cdots, X^n_{-T_h:0}$, each representing a historical motion sequence: (x, y, v_x, v_y, θ). One-hot encoding is associated with respective types. To provide map context for the autonomous (ego) vehicle, we collect map data using a graph search of Lanelets [34], which includes m waypoints $\mathcal{W}_m$ and reference routes $\mathcal{I}$. Each map segment contains coordinate information as well as road features such as road type, speed limit, and the current state of nearby traffic lights.

B. Integrated Prediction and Planning Framework

Fig. 2 illustrates the overall structure of our learning-based framework, which integrates occupancy predictions and planning and consists of three modules. First, we adopt the separate encoding for multi-modal scene inputs $\mathcal{S}$ from our previous work. Encoded visual and vector features are subsequently delivered to occupancy prediction and planner heads. For occupancy heads, we maintain the previous aggregation and decoding pipelines, which then conduct a late fusion of encoded vector and visual features. The prediction heads ultimately output all types of occupancy predictions $\hat{\mathbf{O}}1:T$ for each second of the future horizon. For multi-modal planning heads, we follow the Transformer decoder paradigm and employ a single-layer decoding head for multi-modal path planning $\hat{\mathbf{Y}}^e1:T$. More specifically, we initially introduce a set of k planning anchors $\mathcal{K}_1, \cdot, \mathcal{K}_k$ to guide k modes for planning heads. Each anchor $\mathcal{K}$ can be either static end-point clusters or dynamic ones for learnable embedding that heuristically guide the planning decoding. These anchors then serve as context queries $\mathbf{Q}$ for encoded visual and vector features after self-attentions.

For the multi-head cross-attention module (MHCA), we aim to adaptively obtain interaction-aware features from visual context, vectorized dynamics, and scene topology for the ego vehicle. We implement separate queries using different MHCA blocks for visual and vector features, serving as $\mathbf{K}, \mathbf{V}$. This is due to the length of padded visual features that will sparse the attention scoring when considering concatenated attention heads. After subsequent feed-forward networks (FFNs), each mode will decode to an output of a planning path $\hat{\mathbf{y}}^k_{1:T}$ and corresponding probabilities $\hat{p}_k$ for imitation likelihoods.

We formulate the first-stage learning framework as a multi-task learning paradigm, and the objectives are offline updated jointly using large-scale datasets. For occupancy predictions, as each grid cell forecasts the occupying probabilities under a certain future horizon, we employ focal loss [35] to balance the occupied samples for each type of actor:

$$\mathcal{L}_{pred} = \frac{1}{HWTU} \sum_{x,y,t,u}^{H,W,T,U} \mathrm{FL}(\hat{\mathbf{O}}^u_t(x,y), \mathbf{O}^u_t(x,y)), \quad (2)$$

where H, W represent the occupancy scales and U denotes the actor types. For multi-modal planning, the imitation

学習目標は以下のように更新される：

$$\mathcal{L}_{plan} = \arg\min_{k} \frac{1}{T} \sum_{t}^{T} \mathrm{SL}_1(\hat{y}_t^k - y_t^k) - p_k \log \hat{p}_k, \quad (3)$$

ここで、SL_1は滑らかなL1損失を表し、グランドトゥルースまでの距離が最小となる計画モードは、初期計画経路$\hat{Y}^e_{1:T}$に対して更新される。

C. 予測ガイド付き計画の改良

シーン交通参加者の占有予測$\hat{O}_{1:T}$とマルチモーダル初期計画結果$\hat{Y}^e_{1:T}$が与えられると、OPGPの第2ステージは、計画洗練のための予測ガイドパイプラインを構築する。まず、本節では、予測と計画の両方にFrenet [19]座標を用いた変換パラダイムを設計する。次に、変換された占有予測によって導かれる計画最適化を導入する。

変換：我々は、Frenet座標を変換ターゲットに活用する。これは、最適化計画の難しさを緩和するためである。生成された参照経路Iを仮定すると、各参照点r∈Iは接線ベクトルと法線ベクトルによって曲線フレームで動的に割り当てられる：$[\ \tilde{t}_r, \tilde{n}_r\ ]$。そして、現在の直交座標 $\tilde{y} = (x, y)$ は、次の関係に従って、変換 $F : \tilde{y} \rightarrow \tilde{r}$ を通して Frenet $\tilde{r} = (s, d)$ に変換することができる。

$$\vec{y}(s(t), d(t)) = \vec{r}(s(t)) + d(t)\vec{n}_r(s(t)). \quad (4)$$

初期計画結果$\hat{y}^k_{1:T}$に対して、まずトップスコア軌道$\hat{Y}^e_{1:T} = \arg\max_{pk} \hat{y}^k\ 1 : T$を選択し、$\hat{Y}^{\mathcal{T}}_{1:T} = F(\ \hat{Y}^e_{1:T}\ )$ で変換する。予測された占有率$\hat{O}_{1:T}$に対して、我々の目的は、Iを中心とした占有率予測をフィルタリングし、Frenetフレーム下の占有率にストレッチすることである。この変換は、スケーラブルな占有率の探索空間を、オクルージョンのない計画に焦点を当てる。さらに、Frenet変換された占有率は、計画最適化のための凸性を緩和する。より具体的には、Frenetフレーム上の定義済みメッシュグリッドインデックスが与えられる：I_{sd}, $s \in [0, S]$; $d \in [-\frac{D}{2}, \frac{D}{2}]$, この手続きは、まずFrenetから$F^{-1}$だけデカルトフレームに逆変換する。次に、直交格子から占有格子への関係に従って、画素(w, h)をPとする：

$$(w, h) = \mathrm{int}(\frac{1}{n}(x - x_0^e, y - y_0^e)), \quad (5)$$

ここで、int は丸め関数、n はメートルあたりの画素数を表す。変換された占有予測$\hat{O}^{\mathcal{T}}\ 1 : T$は、次式で集められる：

$$\begin{aligned} \hat{\mathbf{O}}^{\mathbf{T}}_{1:T}(s, d) &= \hat{\mathbf{O}}_{1:T}(\mathbf{I}_w, \mathbf{I}_h), \\ [\mathbf{I}_w; \mathbf{I}_h] &= \mathcal{P}(\mathcal{F}^{-1}(\mathbf{I}_{sd})). \end{aligned} \quad (6)$$

OPGPパイプラインでは、出力変換された計画$\hat{Y}^{\mathcal{T}}_{1:T}$は、計画洗練のために$\hat{O}^{\mathcal{T}}_{1:T}$からの予測のガイダンスで最適化される。

最適化：計画結果の安全性と運動性能を向上させるために、ここでは有限ホライズン下でのオープンループ最適化問題を定式化する。最適化は、変換された占有予測$\hat{O}^{\mathcal{T}}_{1:T}$によって導かれるコスト関数集合Cを最小化する最適なシーケンス$\tau^* = \{\tau_1, - -, \tau_T\}$を探索する。より具体的には、Cはコスト関数$c_i$の集合の重み付き$\omega_i$二乗和である：

$$\begin{aligned} \tau^* &= \arg\min_{\tau} C(\tau, \hat{\mathbf{O}}^{\mathbf{T}}_{1:T}, \mathcal{S}), \\ C &= \sum_i \|\omega_i c_i(\tau^i, \hat{\mathbf{O}}^{\mathbf{T}}_{1:T}, \mathcal{S})\|^2, \tau^i \subseteq \tau. \end{aligned} \quad (7)$$

コスト関数集合Cは、運転進捗、運転快適性、ルールの遵守、そして最も重要な運転安全性など、様々な計画性能を考慮した慎重に設計された重みω_iを持つ様々なコスト項c_iを包含する。コストセットは以下のように明示的にカタログ化されている：

1）運転の進捗：効率的な運転を促進するために、制限速度v^s_{limit}以内の縦方向の速度コストを一定に保ち、より多くのオンロードの進捗を可能にする：

$$\mathbf{c}_t^{\mathrm{progress}} = \dot{s}_t - v^s_{\mathrm{limit}}. \quad (8)$$

2）走行の快適性：計画の快適性と運転の滑らかさを促進するために、加速度とジャークの最小化を導入する：

$$\mathbf{c}_t^{\mathrm{comfort}} = \ddot{s}_t + \ddot{d}_t + \dddot{s}_t. \quad (9)$$

3）ルールの遵守：設計されたADSは交通ルールに適合し、基準車線内に留まることが期待される。そこで、(s, d)の最大マージン目標に基づくオフルートコストと信号機コストをそれぞれ提案する：

$$\mathbf{c}_t^{\mathrm{route}} = d_t. \quad (10)$$

$$\mathbf{c}_t^{\mathrm{tl}} = \begin{cases} \mathbf{1}(s_{\mathrm{red}})(s_t - s_{\mathrm{red}}), & s_t > s_{\mathrm{red}}, \\ 0, & \mathrm{otherwise}. \end{cases} \quad (11)$$

示4）、予測ガイド付き安全性：計画精緻化のための運転
安全性を高めるために、安全な運転のための動的障害物として、変換された占有予測$\hat{O}^{\mathcal{T}}_{1:T}$のガイダンスを活用する。占有率予測は、アクタータイプによって重み付けされた潜在的な衝突コストとみなされる：

$$\mathbf{c}_t^{\mathrm{ogm}}(s, d) = \sum_u^U \lambda_u \hat{\mathbf{O}}^{\mathbf{T}u}_t(s, d), \quad (12)$$

フレーム間の占有率が保たれているという問題から、占有率の予測されるアクターは、長い尾を引いて拡大する可能性がある。ADSの過保守運転につながる可能性がある。そこで、閾値以上の確信衝突コストに対して、確信安全距離s^{safe}_tを設計する：

$$s_t^{\mathrm{safe}} = \min_s(\sum_d^D \mathbf{c}_t^{\mathrm{ogm}}(s, d) > \epsilon), \quad (13)$$

最適化を簡単にするため、計画精密化のための縦断的なガイダンスのみを行う。前方符号距離を考慮すると

$$\mathrm{sgnd}(s_t) = (s_t - \mathbf{I}_s)\mathbf{1}(s_t > \mathbf{I}_s), \quad (14)$$

learning objectives are updated as follows:

$$\mathcal{L}_{plan} = \arg\min_{k} \frac{1}{T} \sum_{t}^{T} \mathrm{SL}_1(\hat{y}_t^k - y_t^k) - p_k \log \hat{p}_k, \quad (3)$$

where SL_1 denotes smooth L1 loss, and the planning modes with minimum distance to ground truth will be updated for the initial planning path $\hat{\mathbf{Y}}_{1:T}^e$.

C. Prediction-guided planning refinement

Given the occupancy prediction for scene traffic participants $\hat{\mathbf{O}}_{1:T}$ as well as the multi-modal initial planning results $\hat{\mathbf{Y}}_{1:T}^e$, the second stage of OPGP constructs a prediction-guided pipeline for planning refinement. First, this section designs a transformation paradigm using Frenet [19] coordinates for both prediction and planning. Then, we introduce a planning optimization guided by the transformed occupancy predictions.

Transformation: We leverage Frenet coordinates for transformation targets, as they relax the difficulties in planning optimizations. Suppose a generated reference route $\mathcal{I}$, each reference point $r \in \mathcal{I}$ is dynamically assigned with a curvilinear frame by tangential and normal vectors: $[\vec{t}_r, \vec{n}_r]$. Then, the current Cartesian coordinate $\vec{y} = (x, y)$ can be transformed into Frenet $\vec{r} = (s, d)$ through the transformation $\mathcal{F} : \vec{y} \rightarrow \vec{r}$ following the relations:

$$\vec{y}(s(t), d(t)) = \vec{r}(s(t)) + d(t)\vec{n}_r(s(t)). \quad (4)$$

For initial planning results $\hat{\mathbf{y}}_{1:T}^k$, we first select the top-scoring trajectories $\hat{\mathbf{Y}}_{1:T}^e = \arg\max_{p_k} \hat{y}^k 1:T$, and then transform them by $\hat{\mathbf{Y}}_{1:T}^{\mathbf{T}} = \mathcal{F}(\hat{\mathbf{Y}}_{1:T}^e)$. For predicted occupancy $\hat{\mathbf{O}}_{1:T}$, our objective is to filter out the occupancy prediction centered on $\mathcal{I}$ and stretch onto occupancy under the Frenet frame. This transformation focuses the searching space of scalable occupancy on planning without occlusions. Moreover, the Frenet-transformed occupancy relieves the convexity for planning optimization. More specifically, given predefined mesh-grid indices on the Frenet frame: $\mathbf{I}_{sd}, s \in [0, S]; d \in [-\frac{D}{2}, \frac{D}{2}]$, this procedure transforms reversely from Frenet firstly onto the Cartesian frame by $\mathcal{F}^{-1}$. Then, following the relations from Cartesian to occupancy grids pixel (w, h) denoted as $\mathcal{P}$:

$$(w, h) = \mathrm{int}(\frac{1}{n}(x - x_0^e, y - y_0^e)), \quad (5)$$

where int is the rounding function and n represents pixels per meter. The transformed occupancy prediction $\hat{\mathbf{O}}^{\mathbf{T}} 1:T$ is then gathered by:

$$\begin{aligned} \hat{\mathbf{O}}_{1:T}^{\mathbf{T}}(s, d) &= \hat{\mathbf{O}}_{1:T}(\mathbf{I}_w, \mathbf{I}_h), \\ [\mathbf{I}_w; \mathbf{I}_h] &= \mathcal{P}(\mathcal{F}^{-1}(\mathbf{I}_{sd})). \end{aligned} \quad (6)$$

In OPGP pipelines, the output transformed planning $\hat{\mathbf{Y}}_{1:T}^{\mathbf{T}}$ will then be optimized with the guidance of predictions from $\hat{\mathbf{O}}_{1:T}^{\mathbf{T}}$ for planning refinement.

Optimization: To ensure enhanced safety and motion performance for planning results, here we formulate an open-loop optimization problem under finite horizons. The optimization searches for an optimal sequence $\tau^* = \{\tau_1, \cdots, \tau_T\}$ that minimizes the cost function sets C guided by transformed occupancy predictions $\hat{\mathbf{O}}_{1:T}^{\mathbf{T}}$. More specifically, C is the weighted ω_i squared sum of a set of cost functions c_i:

$$\begin{aligned} \tau^* &= \arg\min_{\tau} C(\tau, \hat{\mathbf{O}}_{1:T}^{\mathbf{T}}, \mathcal{S}), \\ C &= \sum_i \|\omega_i c_i(\tau^i, \hat{\mathbf{O}}_{1:T}^{\mathbf{T}}, \mathcal{S})\|^2, \tau^i \subseteq \tau. \end{aligned} \quad (7)$$

The cost function set C embraces various cost terms c_i with carefully-devised weights ω_i that consider a variety of planning performances, including driving progress, driving comfort, adherence to rules, and most importantly, driving safety. The cost set is explicitly cataloged as follows:

1) Driving progress: To encourage efficient driving, we assist a consistent longitudinal speed cost within speed limits v_{limit}^s, allowing more on-road progress:

$$\mathbf{c}_t^{\mathrm{progress}} = \dot{s}_t - v_{\mathrm{limit}}^s. \quad (8)$$

2) Driving comfort: To facilitate planning comfort and driving smoothness, we introduce the minimization for accelerations and jerks:

$$\mathbf{c}_t^{\mathrm{comfort}} = \ddot{s}_t + \ddot{d}_t + \dddot{s}_t. \quad (9)$$

3) Adherence to rules: We expect the designed ADS to conform to traffic rules and stay within reference lanes. Thus, we propose an off-route cost and a traffic light cost based on max-margin objectives on (s, d) respectively:

$$\mathbf{c}_t^{\mathrm{route}} = d_t. \quad (10)$$

$$\mathbf{c}_t^{\mathrm{tl}} = \begin{cases} \mathbf{1}(s_{\mathrm{red}})(s_t - s_{\mathrm{red}}), & s_t > s_{\mathrm{red}}, \\ 0, & \mathrm{otherwise}\,. \end{cases} \quad (11)$$

where s_{red} is the stop point for red-light reference lane.

4) Prediction guided safety: To enhance driving safety for planning refinement, we leverage the guidance of transformed occupancy predictions $\hat{\mathbf{O}}_{1:T}^{\mathbf{T}}$ as dynamic obstacles for safe driving. The occupancy prediction is considered as a potential collision cost weighted by actor types:

$$\mathbf{c}_t^{\mathrm{ogm}}(s, d) = \sum_u^U \lambda_u \hat{\mathbf{O}}_t^{\mathbf{T}u}(s, d), \quad (12)$$

Due to the issues of preserved occupancy between frames, the predicted actors in occupancy are possibly enlarged with long tails. This may lead to over-conservative driving for ADS. Therefore, we design a confidence safety distance s_t^{safe} for the confident collision cost over threshold ϵ:

$$s_t^{\mathrm{safe}} = \min_s (\sum_d^D \mathbf{c}_t^{\mathrm{ogm}}(s, d) > \epsilon), \quad (13)$$

For optimization simplicity, we only conduct longitudinal guidance for planning refinement. Considering the forward sign distance:

$$\mathrm{sgnd}(s_t) = (s_t - \mathbf{I}_s)\mathbf{1}(s_t > \mathbf{I}_s), \quad (14)$$

予測ガイド付き安全コストは、符号距離がs_t^{safe}を超える場合の加重予測衝突コスト和としてモデル化される:

$$\mathbf{c}_t^{\text{safe}} = \sum_s^S \sum_d^D \text{sgnd}(s_t)\mathbf{c}_t^{\text{ogm}}(s,d)\mathbf{1}(s > s_t^{\text{safe}}). \quad (15)$$

この非線形最適化問題を解くには、式(4)のようになる。7. $\hat{Y}_{1:T}^{T}$を初期値として計画変数τを繰り返し精緻化するガウス・ニュートン法[37]を採用する。

IV. 実験

A. 実験セットアップ

我々はWaymo Open Motion Dataset (WOMD) [38]を利用する。WOMDは50万以上のサンプルを含み、実世界の幅広い運転シナリオをカバーしている。車両、自転車、歩行者など、様々なタイプの交通エージェント間の複雑な相互作用とダイナミクスを捉えることができる。第1段階でのシーンコンテキストSを入力として、各アクターの履歴状態を過去1秒間10Hzでサンプリングし(T_h = 10)、予測と計画の目的は将来5秒間で予測に1Hz、計画に10Hzである(T = 50)。入出力のBEV画像解像度はH, W = 128で、実世界ではn = 1.6ピクセル/mをレンダリングする。隠れ次元を96とする。ベクトル入力の場合、AVまでの平均距離でソートされた最大n_x = 31の周囲アクターを維持する。標準的な尺度に従い、マルチモーダル計画復号化におけるアンカーk = 6を設定する。OPGPの第2ステージでは、参照ルートIは0.1mごとに滑らかに交差し、変換された占有予測$\hat{O}_t^T$のサイズはS = 1000, D = 20である。計画のために、記録された20秒間のWOMDをランダムに分割し、トレーニング用に297,669サンプル、検証とテスト用に関心のあるシナリオを選択した(各47,728)。

メトリクス OPGPの性能を総合的に評価するため、占有率予測には、チャレンジ[9]で提案された標準的なメトリクスを採用した。予測指標には$\hat{O}_{1:T}$のAUCとsIOUが含まれる。計画指標としては、人間らしさ、運動能力、テスト中の安全性に関する運転性能に注目する。

ベースライン OPGPフレームワークの予測性能と計画性能を検証するために、いくつかの強力なベースラインと比較する。1) 占有予測のために、プランニングヘッドを持たないバニラSTrajNet [18]と、Swin Transformerボトルネック[39]を用いた簡略化された集約とエンコーディングを持つ改良されたベースライン予測器であるVectorFlow-SwinTを提供する。2) オープンループ計画については、OPGPの第1段のみであるVanilla ILや、OPGPのボトルネックで考案したDIM[36]と比較する。これらのベースラインは、OPGPで提案されたモジュールをいくつか用いないアブレーション手法としても機能する。

B. 実装の詳細

GELU活性化関数を選択し、オーバーフィッティングの問題を軽減するために、各層の後に0.1のドロップアウト率を適用する。データ量が多いため、4つのNVIDIA A100 GPUで分散戦略を採用し、総バッチサイズは64である。AdamWオプティマイザを初期学習率1e-4で使用し、コサインアニーリング学習率戦略を採用した。学習エポックの総数を30とする。

C. 定量的結果

1) 予測性能 OPGPのテスト性能を他の強力なベースラインと比較して報告する。表IIに示すように、占有予測タスクのために特別に設計されたバニラSTrajNetと比較すると、OPGPの予測性能はわずかに低下している。これは、OPGPが予測ヘッドと計画ヘッドの両方を持つマルチタスク学習パイプラインであるためである。さらに重要なことは、現在の設計では、予測から計画への特徴ガイダンスのみを考慮していることである。その結果、マルチモーダル計画の学習目的は、占有率予測のためのデコーディングヘッドへの更なる貢献が必要となる。

2) 計画性能: オープンループ計画テストを実施する。表Iに示すように、OPGPの第1段を採用したVanilla ILは、最も低い計画誤差を達成し、k個のアンカーを用いたマルチモーダル構成の性能向上を実証している。一方、DIMの模倣計画のためのユニモーダルガウシアンは、複雑なシナリオの不確実性を解決できない。提案するOPGPは、ILと比較して安全性を高めながら、低い計画誤差をもたらす。第2ステージの占有率予測ガイド付き最適化により、衝突率とオフルート率は大幅に改善される。よく設計されたコスト関数は、モーションメトリクスのジャークを減少させることで、運転の快適性を促進する。

D. Qualitative Results

OPGPの有効性をより深く理解するために、まず、占有率予測の可視化を採用し、テスト中の複数の代表的な運転シナリオにおける性能を評価する。図3に示すように、提案手法は、観測されたアクター(a, c, d)と高オクルージョンリスクシナリオ(a, c)の両方を効果的に処理する。予測精度は車両に対してよく維持されているが、現在の手法では、より小さなアクター(b, d)に対してのみ3sの予測性能を確保することができる。計画のための予測誘導能力を実証するために、計画結果と対応する変換された占有予測を表示する。

図4に示すように、シアン色で表示されたプランニングは、予測された歩行者の接近により、AV(赤)を停止させる。さらに、赤い点線円で示すように、OPGPにおける占有率予測は、エージェント単位の手法では見逃されるような未検出のアクターを予測することができる。

E. 限界と今後の課題

現在の2段階OPGPは有望な結果を示しているが、さらなる改善が必要なトピックもいくつかある。最初の問題は、マルチタスク学習における占有予測性能のわずかな低下である。ネットワーク設計から学習スキームへの共同インタラクションが必要である。

The prediction-guided safety cost is modeled as a weighted predicted collision cost sum for sign distance that surpasses s_t^{safe}:

$$\mathbf{c}_t^{\text{safe}} = \sum_{s}^{S} \sum_{d}^{D} \text{sgnd}(s_t) \mathbf{c}_t^{\text{ogm}}(s,d) \mathbf{1}(s > s_t^{\text{safe}}). \quad (15)$$

To solve this non-linear optimization problem, as depicted in Equ. 7. We employ Gauss-Newton method [37] that iterative refine the planning variable τ based on $\hat{\mathbf{Y}}_{1:T}^{\mathbf{T}}$ as initial value.

IV. Experiments

A. Experimental Setup

We utilize the Waymo Open Motion Dataset (WOMD) [38], which contains over 500,000 samples that cover a wide range of real-world driving scenarios. It captures the complex interactions and dynamics among various types of traffic agents, including vehicles, cyclists, and pedestrians. For scene context $\mathcal{S}$ as inputs at the first stage, the historical states for each actor are sampled at 10Hz for the past one second ($T_h = 10$), while the prediction and planning objectives are over the future 5 seconds at 1Hz for prediction and 10Hz for planning ($T = 50$). The BEV image resolution for input and output is $H, W = 128$, rendering $n = 1.6$ pixels per meter in the real world. We set the hidden dimension as 96. For vector inputs, we maintain a maximum of $n_x = 31$ surrounding actors sorted by their average distance to the AV. Following the standard measure, we set the anchors $k = 6$ in multi-modal planning decoding. For the second stage of OPGP, the reference route $\mathcal{I}$ is smoothly intersected every 0.1m, and the size of transformed occupancy prediction $\hat{\mathbf{O}}_t^{\mathbf{T}}$ is $S = 1000$, $D = 20$. For planning purposes, we randomly split the recorded 20 seconds of WOMD, resulting in 297,669 samples for training and selected scenarios of interest for validation and testing (47,728 each).

Metrics: To ensure a comprehensive evaluation of OPGP performance, for occupancy predictions, we adopted the standard metrics proposed in challenges [9]. The prediction metrics include **AUC** and **sIOU** for $\hat{\mathbf{O}}_{1:T}$. For planning metrics, we focus on the driving performance concerning human-likeness, motion capabilities, and safety during testing.

Baselines: To validate the prediction and planning performance of the OPGP framework, we compare it with several strong baselines. **1) For occupancy prediction**, we provide the vanilla STrajNet [18] without a planning head and VectorFlow-SwinT, the improved baseline predictor with simplified aggregation and encoding using the Swin Transformer bottleneck [39]. **2) For open-loop planning**, we compare it with Vanilla IL, i.e., only the first stage of OPGP, and DIM [36] devised with OPGP bottlenecks. These baselines also serve as ablative methods without some of the proposed modules in OPGP.

B. Implementation Details

We select the GELU activation function and apply a dropout rate of 0.1 after each layer to mitigate overfitting issues. Due to the large volume of data, we employ a distributed strategy on four NVIDIA A100 GPUs with a total batch size of 64. The AdamW optimizer is used with an initial learning rate of 1e-4, and a cosine annealing learning rate strategy is employed. The total number of training epochs is set to 30.

C. Quantitative Results

1) Prediction Performance: We report the testing performance of OPGP in comparison with other strong baselines. As shown in Table II, when compared with the Vanilla STrajNet, which is specifically designed for the occupancy prediction task, the prediction performance of OPGP exhibits a slight decrease. This is due to OPGP being a multi-task learning pipeline with both prediction and planning heads. More importantly, the current design only considers feature guidance from prediction towards planning. Consequently, the training objectives for multi-modal planning require further contributions to the decoding head for occupancy prediction.

2) Planning Performance: We conduct open-loop planning testing. As shown in Table I, Vanilla IL, which employs the first stage of OPGP, achieves the lowest planning errors, demonstrating the performance gain of multi-modal configurations using k anchors. In contrast, a uni-modal Gaussian for imitative planning in DIM cannot resolve the uncertainty in complex scenarios. The proposed OPGP yields low planning errors while enhancing safety compared to IL. The collision and off-route rates are significantly improved due to the occupancy prediction-guided optimization in the second stage. The well-designed cost function also promotes driving comfort by reducing jerks in motion metrics.

D. Qualitative Results

To gain a better understanding of the effectiveness of our OPGP, we first employ visualizations of occupancy prediction to assess its performance on multiple representative driving scenarios during testing. As shown in Fig. 3, the proposed method effectively handles both observed actors (a, c, d) and high-occlusion risk scenarios (a, c). Prediction accuracy is well-maintained for vehicles, but current methods can only ensure 3s prediction performance for smaller actors (b, d). To demonstrate the prediction-guided capability for planning, we display the planning result with corresponding transformed occupancy predictions.

As illustrated in Fig. 4, the planning rendered in cyan causes the AV (red) to stop due to the predicted approaching pedestrians. Moreover, as indicated by the red dotted circle, the occupancy prediction in OPGP can predict undetected actors that would be missed by agent-wise methods.

E. Limitations and Future Work

Although the current two-stage OPGP shows promising results, it also raises some topics that require further improvements. The first issue is the slight drop in occupancy prediction performance in multi-task learning. This requires joint interactions from network design to the learning scheme

TABLE I

(WAYMOモーションデータセットにおけるベンループテスト結果

	Safety (%)			Motion (ms^{-2}, ms^{-3})			Planning errors (m)		
Models	Collisions	Off route	Red light	Jerk	Acc.	Lat.Acc.	1s	3s	5s
Vanilla-IL	7.575	2.76	1.671	4.323	0.676	0.129	**0.195**	**1.076**	**2.759**
DIM-OPGP [36]	9.275	14.27	3.246	7.226	0.977	0.252	0.407	1.865	4.473
Vanilla-Test	-	-	-	4.02	0.692	0.167	-	-	-
OPGP	**3.018**	**1.809**	**0.501**	3.869	0.551	0.156	0.226	1.21	3.184

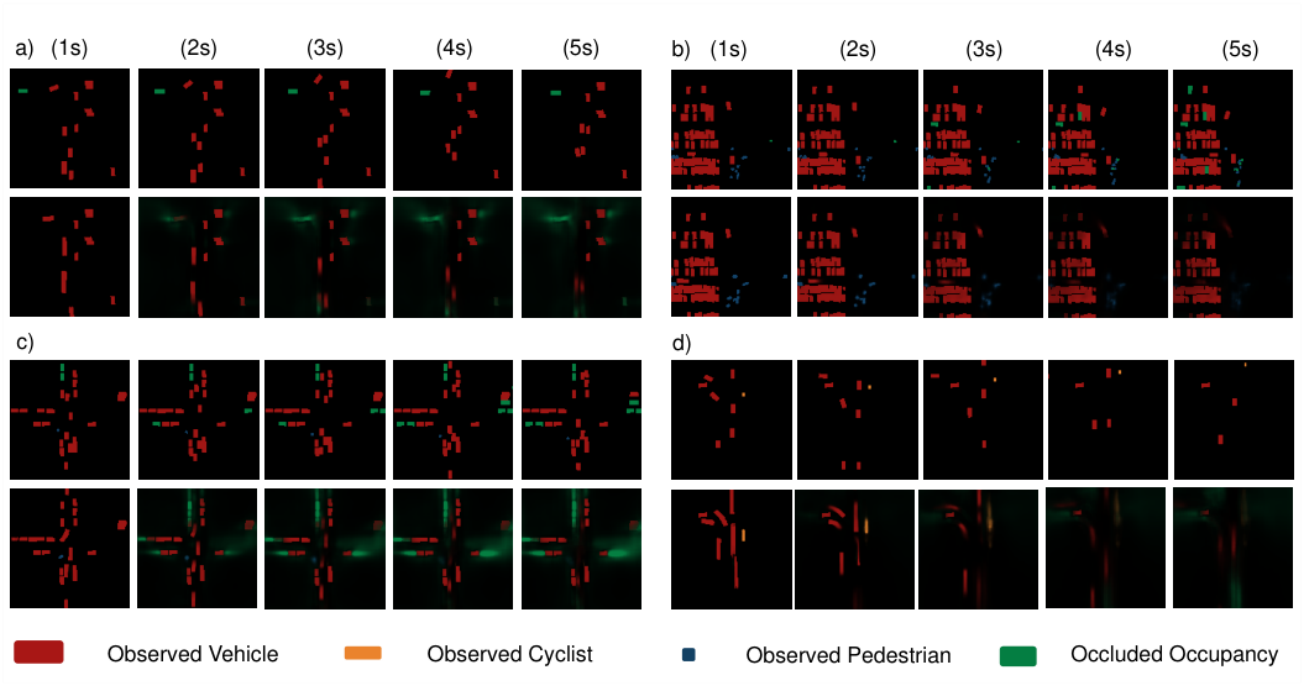

図3. テストセットからの占有率予測の定性的結果。この図には、a)合流を伴うT字路、b)多数の歩行者がいる駐車場、c)閉塞した交差点、d)隣接する自転車とのフォークロード、といったいくつかの運転シナリオが示されている。

TABLE II

T占有率予測に関するテスト結果

Models	Vec-AUC	Vec-sIOU	Occ-AUC	Occ-sIOU
STrajNet [18]	**0.856**	**0.696**	**0.146**	**0.023**
VectorFlow [39]	0.813	0.656	0.112	0.017
OPGP	0.854	0.679	0.128	0.02

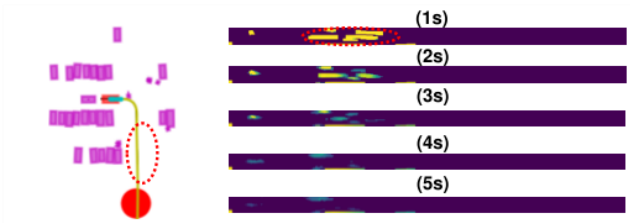

図4. 図4. ガイダンスに対応する変換された占有率予測による洗練された計画結果の例。歩行者に接近したため、AV(赤)が停止する。赤い点で囲まれた予測結果は、エージェント中心の手法における未検出のアクターを指す。

を予測・計画に利用する。また、手作業による重み付けや保守的な予測ガイダンスのチューニングにコストがかかるため、手動で安全距離を定義する必要がある。占有率予測は、難解なピクセルを除去するための追加ガイダンスを必要とする。

features:

今後の課題としては、OPGPパイプラインに基づくネットワーク構造の改善により、これらの問題を解決することである。

V. Conclusions

本論文では、予測ガイダンスによって学習ベースのプランニングを共同で改良する2段階の占有予測誘導型ニューラルプランナー(OPGP)を紹介する。我々は、包括的な占有予測とマルチモーダルな計画目標のために設計された、Transformerバックボーンを持つ統合学習ベースのフレームワークを開発する。予測と計画のための第一段階の出力に続いて、曲線フレーム上に構築された変換された占有ガイド付き最適化は、手作りのコスト関数設計の使用を通じて、直接的な計画の洗練を達成する。予測性能と計画性能は、大規模な実世界データセット(WOMD)を用いて広範囲に検証されている。バニラの強力な予測ベースラインと比較して、ロバストな性能を示すプランニング結果は、安全性と運転の滑らかさの向上を示している。さらに、定性的な結果は、変換された占有予測ガイダンスの有効性を立証し、エージェント単位の方法と比較して、検出されず、隠蔽されたアクターを扱う際のスケーラビリティの向上を明らかにする。

TABLE I

OPEN-LOOP TESTING RESULTS ON WAYMO MOTION DATASET

	Safety (%)			Motion (ms^{-2}, ms^{-3})			Planning errors (m)		
Models	Collisions	Off route	Red light	Jerk	Acc.	Lat.Acc.	1s	3s	5s
Vanilla-IL	7.575	2.76	1.671	4.323	0.676	0.129	**0.195**	**1.076**	**2.759**
DIM-OPGP [36]	9.275	14.27	3.246	7.226	0.977	0.252	0.407	1.865	4.473
Vanilla-Test	-	-	-	4.02	0.692	0.167	-	-	-
OPGP	**3.018**	**1.809**	**0.501**	3.869	0.551	0.156	0.226	1.21	3.184

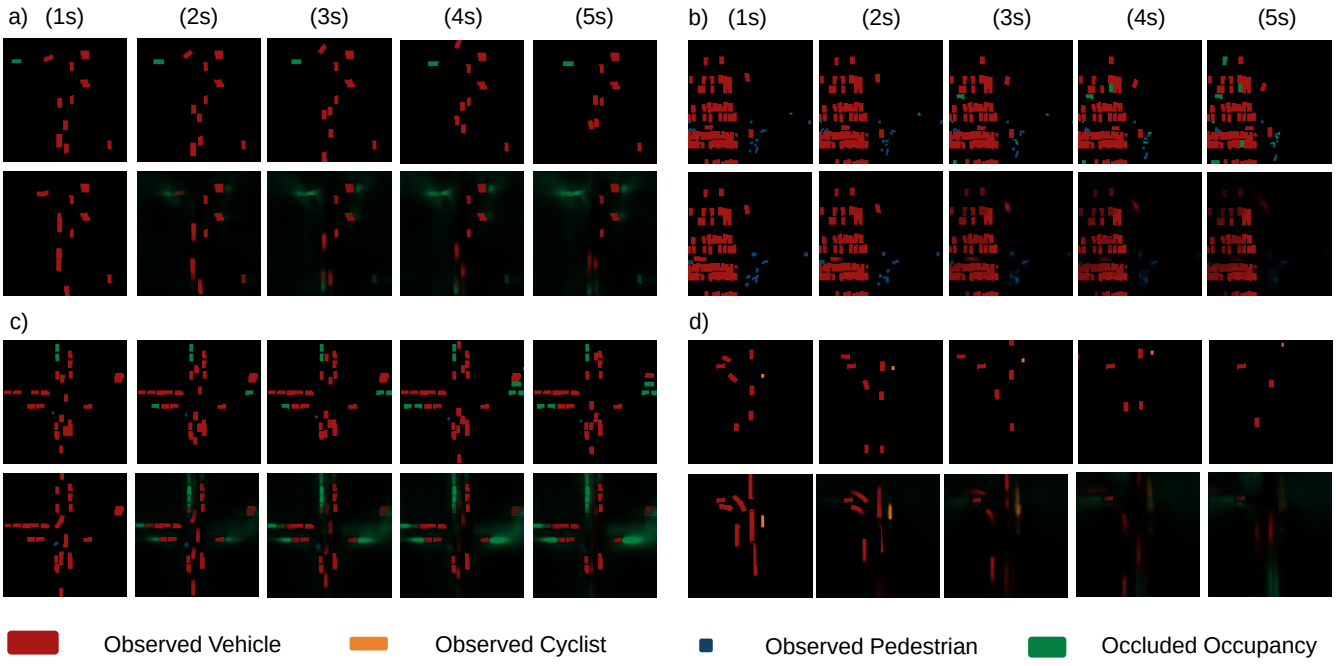

Fig. 3. Qualitative results of occupancy predictions from the testing set. The sub-figures display several selected driving scenarios with specific features: a) T-intersections with merging; b) Parking lots with numerous pedestrians; c) Occluded intersections; d) Fork roads with adjacent cyclists.

TABLE II

TESTING RESULTS ON OCCUPANCY PREDICTIONS

Models	Vec-AUC	Vec-sIOU	Occ-AUC	Occ-sIOU
STrajNet [18]	**0.856**	**0.696**	**0.146**	**0.023**
VectorFlow [39]	0.813	0.656	0.112	0.017
OPGP	0.854	0.679	0.128	0.02

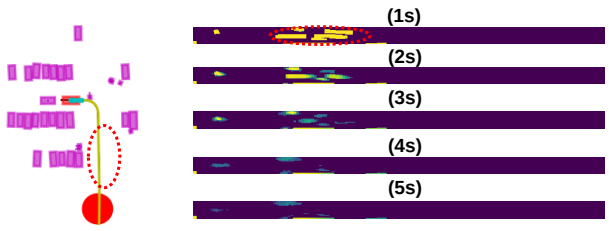

Fig. 4. An example of refined planning results with corresponding transformed occupancy prediction for guidance. The AV (red) stops due to the approaching pedestrians. The prediction results circled in red dots refer to undetected actors in agent-centric methods.

for prediction and planning. Another issue is the costly tuning for handcrafted weights and conservative prediction guidance, which forces us to manually define a safety distance. Occupancy prediction requires additional guidance in eliminating intractable pixels. Future work will aim to resolve these issues by improving network structures based on OPGP pipelines.

V. CONCLUSIONS

In this paper, we present a two-stage Occupancy Prediction-Guided Neural Planner (OPGP) that refines learning-based planning through prediction guidance in a joint manner. We develop an integrated learning-based framework with Transformer backbones, designed for comprehensive occupancy predictions and multi-modal planning objectives. Following the first stage outputs for prediction and planning, a transformed occupancy-guided optimization, built upon a curvilinear frame, achieves direct planning refinement through the use of handcrafted cost function designs. The prediction and planning performance are extensively validated using large-scale, real-world datasets (WOMD). Exhibiting robust performance in comparison to vanilla strong prediction baselines, the planning results demonstrate enhanced safety and driving smoothness. Furthermore, qualitative results substantiate the effectiveness of the transformed occupancy prediction guidance, revealing increased scalability in handling undetected and occluded actors when compared to agent-wise methods.

REFERENCES

[1] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2020.

[2] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li *et al.*, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, 2022.

[3] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz *et al.*, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, 2021.

[4] W. Wang, L. Wang, C. Zhang, C. Liu, L. Sun *et al.*, "Social interactions for autonomous driving: A review and perspectives," *Foundations and Trends® in Robotics*, vol. 10, no. 3-4, pp. 198–376, 2022.

[5] X. Mo, Y. Xing, and C. Lv, "Interaction-aware trajectory prediction of connected vehicles using cnn-lstm networks," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2020, pp. 5057–5062.

[6] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 303–15 312.

[7] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 683–700.

[8] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," *arXiv preprint arXiv:2303.05760*, 2023.

[9] R. Mahjourian, J. Kim, Y. Chai, M. Tan, B. Sapp, and D. Anguelov, "Occupancy flow fields for motion forecasting in autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5639–5646, 2022.

[10] J. Kim, R. Mahjourian, S. Ettinger, M. Bansal, B. White, B. Sapp, and D. Anguelov, "Stopnet: Scalable trajectory and occupancy prediction for urban autonomous driving," *arXiv preprint arXiv:2206.00991*, 2022.

[11] H. Zhou, J. Laval, A. Zhou, Y. Wang, W. Wu, Z. Qing, and S. Peeta, "Review of learning-based longitudinal motion planning for autonomous vehicles: research gaps between self-driving and traffic congestion," *Transportation research record*, vol. 2676, no. 1, pp. 324–341, 2022.

[12] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 740–759, 2020.

[13] H. Liu, Z. Huang, X. Mo, and C. Lv, "Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving," *arXiv preprint arXiv:2208.12263*, 2022.

[14] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal *et al.*, "Scene transformer: A unified architecture for predicting multiple agent trajectories," *arXiv preprint arXiv:2106.08417*, 2021.

[15] J. L. V. Espinoza, A. Liniger, W. Schwarting, D. Rus, and L. Van Gool, "Deep interactive motion prediction and planning: Playing games with motion prediction models," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 1006–1019.

[16] Z. Huang, H. Liu, J. Wu, and C. Lv, "Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving," *arXiv preprint arXiv:2207.10422*, 2022.

[17] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "Lookout: Diverse multi-future prediction and planning for self-driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 107–16 116.

[18] H. Liu, Z. Huang, and C. Lv, "Strajnet: Occupancy flow prediction via multi-modal swin transformer," *arXiv preprint arXiv:2208.00394*, 2022.

[19] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a frenet frame," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 987–993.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[21] Q. Sun, X. Huang, J. Gu, B. C. Williams, and H. Zhao, "M2i: From factored marginal trajectory prediction to interactive prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6543–6552.

[22] X. Mo, Z. Huang, Y. Xing, and C. Lv, "Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[23] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "Gohome: Graph-oriented heatmap output for future motion estimation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9107–9114.

[24] ——, "Thomas: Trajectory heatmap output with learned multi-agent sampling," *arXiv preprint arXiv:2110.06607*, 2021.

[25] M. Bansal, A. Krizhevsky, and A. Ogale, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst," *arXiv preprint arXiv:1812.03079*, 2018.

[26] P. Hang, C. Lv, C. Huang, J. Cai, Z. Hu, and Y. Xing, "An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors," *IEEE transactions on vehicular technology*, vol. 69, no. 12, pp. 14 458–14 469, 2020.

[27] Z. Huang, J. Wu, and C. Lv, "Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[28] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *IEEE Transactions on intelligent transportation systems*, vol. 17, no. 4, pp. 1135–1145, 2015.

[29] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "Pip: Planning-informed trajectory prediction for autonomous driving," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 598–614.

[30] A. Kamenev, L. Wang, O. B. Bohan, I. Kulkarni, B. Kartal, A. Molchanov, S. Birchfield, D. Nistér, and N. Smolyanskiy, "Predictionnet: Real-time joint probabilistic traffic prediction for planning, control, and simulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8936–8942.

[31] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 403–14 412.

[32] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*. PMLR, 2023, pp. 726–737.

[33] Z. Huang, X. Mo, and C. Lv, "Recoat: A deep learning-based framework for multi-modal motion prediction in autonomous driving application," *arXiv preprint arXiv:2207.00726*, 2022.

[34] P. Bender, J. Ziegler, and C. Stiller, "Lanelets: Efficient map representation for autonomous driving," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 420–425.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[36] N. Rhinehart, R. McAllister, and S. Levine, "Deep imitative models for flexible inference, planning, and control," *arXiv preprint arXiv:1810.06544*, 2018.

[37] M. Bhardwaj, B. Boots, and M. Mukadam, "Differentiable gaussian process motion planning," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 10 598–10 604.

[38] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[39] X. Huang, X. Tian, J. Gu, Q. Sun, and H. Zhao, "Vectorflow: Combining images and vectors for traffic occupancy and flow prediction," *arXiv preprint arXiv:2208.04530*, 2022.

REFERENCES

[1] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2020.

[2] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li *et al.*, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, 2022.

[3] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz *et al.*, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, 2021.

[4] W. Wang, L. Wang, C. Zhang, C. Liu, L. Sun *et al.*, "Social interactions for autonomous driving: A review and perspectives," *Foundations and Trends® in Robotics*, vol. 10, no. 3-4, pp. 198–376, 2022.

[5] X. Mo, Y. Xing, and C. Lv, "Interaction-aware trajectory prediction of connected vehicles using cnn-lstm networks," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2020, pp. 5057–5062.

[6] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 303–15 312.

[7] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 683–700.

[8] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," *arXiv preprint arXiv:2303.05760*, 2023.

[9] R. Mahjourian, J. Kim, Y. Chai, M. Tan, B. Sapp, and D. Anguelov, "Occupancy flow fields for motion forecasting in autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5639–5646, 2022.

[10] J. Kim, R. Mahjourian, S. Ettinger, M. Bansal, B. White, B. Sapp, and D. Anguelov, "Stopnet: Scalable trajectory and occupancy prediction for urban autonomous driving," *arXiv preprint arXiv:2206.00991*, 2022.

[11] H. Zhou, J. Laval, A. Zhou, Y. Wang, W. Wu, Z. Qing, and S. Peeta, "Review of learning-based longitudinal motion planning for autonomous vehicles: research gaps between self-driving and traffic congestion," *Transportation research record*, vol. 2676, no. 1, pp. 324–341, 2022.

[12] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 740–759, 2020.

[13] H. Liu, Z. Huang, X. Mo, and C. Lv, "Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving," *arXiv preprint arXiv:2208.12263*, 2022.

[14] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal *et al.*, "Scene transformer: A unified architecture for predicting multiple agent trajectories," *arXiv preprint arXiv:2106.08417*, 2021.

[15] J. L. V. Espinoza, A. Liniger, W. Schwarting, D. Rus, and L. Van Gool, "Deep interactive motion prediction and planning: Playing games with motion prediction models," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 1006–1019.

[16] Z. Huang, H. Liu, J. Wu, and C. Lv, "Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving," *arXiv preprint arXiv:2207.10422*, 2022.

[17] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "Lookout: Diverse multi-future prediction and planning for self-driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 107–16 116.

[18] H. Liu, Z. Huang, and C. Lv, "Strajnet: Occupancy flow prediction via multi-modal swin transformer," *arXiv preprint arXiv:2208.00394*, 2022.

[19] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a frenet frame," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 987–993.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[21] Q. Sun, X. Huang, J. Gu, B. C. Williams, and H. Zhao, "M2i: From factored marginal trajectory prediction to interactive prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6543–6552.

[22] X. Mo, Z. Huang, Y. Xing, and C. Lv, "Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[23] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "Gohome: Graph-oriented heatmap output for future motion estimation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9107–9114.

[24] ——, "Thomas: Trajectory heatmap output with learned multi-agent sampling," *arXiv preprint arXiv:2110.06607*, 2021.

[25] M. Bansal, A. Krizhevsky, and A. Ogale, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst," *arXiv preprint arXiv:1812.03079*, 2018.

[26] P. Hang, C. Lv, C. Huang, J. Cai, Z. Hu, and Y. Xing, "An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors," *IEEE transactions on vehicular technology*, vol. 69, no. 12, pp. 14 458–14 469, 2020.

[27] Z. Huang, J. Wu, and C. Lv, "Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[28] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *IEEE Transactions on intelligent transportation systems*, vol. 17, no. 4, pp. 1135–1145, 2015.

[29] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "Pip: Planning-informed trajectory prediction for autonomous driving," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 598–614.

[30] A. Kamenev, L. Wang, O. B. Bohan, I. Kulkarni, B. Kartal, A. Molchanov, S. Birchfield, D. Nistér, and N. Smolyanskiy, "Predictionnet: Real-time joint probabilistic traffic prediction for planning, control, and simulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8936–8942.

[31] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 403–14 412.

[32] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*. PMLR, 2023, pp. 726–737.

[33] Z. Huang, X. Mo, and C. Lv, "Recoat: A deep learning-based framework for multi-modal motion prediction in autonomous driving application," *arXiv preprint arXiv:2207.00726*, 2022.

[34] P. Bender, J. Ziegler, and C. Stiller, "Lanelets: Efficient map representation for autonomous driving," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 420–425.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[36] N. Rhinehart, R. McAllister, and S. Levine, "Deep imitative models for flexible inference, planning, and control," *arXiv preprint arXiv:1810.06544*, 2018.

[37] M. Bhardwaj, B. Boots, and M. Mukadam, "Differentiable gaussian process motion planning," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 10 598–10 604.

[38] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[39] X. Huang, X. Tian, J. Gu, Q. Sun, and H. Zhao, "Vectorflow: Combining images and vectors for traffic occupancy and flow prediction," *arXiv preprint arXiv:2208.04530*, 2022.