SparseDrive: スパースシーン表現によるエンドツーエンドの自律走行

Wenchao Sun[1,2] **Xuewu Lin**[2] **Yining Shi**[1] **Chuang Zhang**[1] **Haoran Wu**[1] **Sifa Zheng**[1]

[1] Tsinghua University [2] Horizon

Abstract

確立されたモジュール型自律走行システムは、知覚、予測、計画など、モジュール間の情報損失やエラー蓄積に悩まされる、異なる独立したタスクに分離される。対照的に、エンドツーエンドのパラダイムは、マルチタスクを完全に微分可能なフレームワークに統一し、計画指向の精神で最適化を可能にする。エンドツーエンドのパラダイムは大きな可能性を秘めているにもかかわらず、既存の手法の性能と効率は、特に計画の安全性の点で満足できるものではない。これは、計算コストの高いBEV(鳥瞰図)機能と、予測・計画のための素直な設計によるものである。この目的のために、我々はスパース表現を探求し、エンドツーエンドの自律走行のためのタスク設計をレビューし、SparseDriveと名付けられた新しいパラダイムを提案する。具体的には、SparseDriveは対称スパース知覚モジュールと並列モーションプランナから構成される。スパース知覚モジュールは、検出、追跡、オンラインマッピングを対称モデルアーキテクチャで統合し、走行シーンの完全スパース表現を学習する。動き予測とプランニングについては、これら2つのタスクの大きな類似性をレビューし、モーションプランナーの並列設計を導く。この並列設計に基づき、計画をマルチモーダル問題としてモデル化し、合理的で安全な軌道を最終的な計画出力として選択するために、衝突を考慮した再スコアモジュールを組み込んだ階層的な計画選択戦略を提案する。このような効果的な設計により、SparseDriveは全てのタスクの性能において、従来の最先端技術を大差で上回り、より高い学習効率と推論効率を達成した。コードは https://github.com/swc-17/SparseDrive で検証可能である。

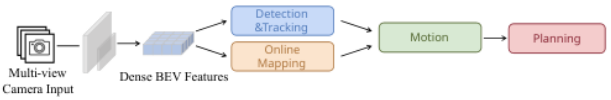
1 Introduction

従来の自律走行システムは、順次モジュール化されたタスクとして特徴付けられる。解釈とエラー追跡には有利であるが、必然的に連続するモジュール間で情報損失と累積エラーが生じ、システムの最適な性能ポテンシャルが制限される。

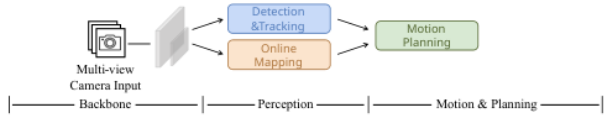
最近、エンドツーエンドの運転パラダイムが有望な研究方向として浮上した。このパラダイムは、すべてのタスクを1つの全体的なモデルに統合し、最終的なプランニングの追求に向けて最適化することができる。しかし、既存の手法[15, 20]は、性能と効率の点で満足できるものではない。一方、従来の手法は、計算コストの高いBEV特徴量に依存している。一方、予測とプランニングのためのシンプルな設計は、モデルの性能を制限する。図1aにBEV-Centricパラダイムとしての先行手法をまとめる。

エンドツーエンドパラダイムの可能性を十分に活用するために、既存の手法のタスク設計をレビューし、以下のように、動作予測とプランニングの間に共有される3つの主要な類似性が無視されることを主張する：(1) 周囲のエージェントと自車両の将来の軌道を予測することを目的とする。

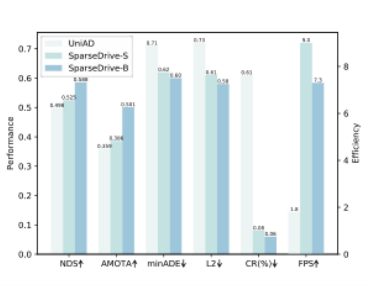
Preprint. Under review.



(a) BEV中心パラダイム。



(b) スパース中心パラダイム。



(c) 本手法とプライベートSOTA[15]の比較。

図1:様々なエンドツーエンドパラダイムの比較。(a) BEV-Centricパラダイム。(b) 提案するスパース中心パラダイム。(c) (a)と(b)の性能と効率の比較。

モーション予測とプランニングは、道路エージェント間の高次かつ双方向の相互作用を考慮する必要がある。しかし、従来の手法では、一般的に動きの予測と計画に逐次設計を採用しており、自車両が周囲のエージェントに与える影響を無視している。(2) 将来の軌道を正確に予測するためには、シーン理解のための意味情報と、エージェントの将来の動きを予測するための幾何学的情報が必要であり、これは動き予測とプランニングの両方に適用可能である。これらの情報は、周囲のエージェントの上流知覚タスクでは抽出されるが、エゴ・ビークルでは見落とされる。(3)運動予測もプランニングも不確実性を伴うマルチモーダル問題であるが、従来の手法ではプランニングのための決定論的な軌道しか予測できなかった。

このため、図1bに示すようなスパース中心パラダイムであるSparseDriveを提案する。具体的には、SparseDriveは対称スパース知覚モジュールと並列モーションプランナから構成される。非結合インスタンス特徴と幾何学的アンカーを1つのインスタンス(動的道路エージェントまたは静的マップ要素)の完全な表現として、対称スパース知覚は、対称モデルアーキテクチャで検出、追跡、オンラインマッピングタスクを統合し、完全にスパースなシーン表現を学習する。並列モーションプランナーでは、まず、エゴインスタンス初期化モジュールから、意味と幾何学を考慮したエゴインスタンスを得る。疎な知覚からのエゴインスタンスと周囲のエージェントインスタンスを用いて、動き予測とプランニングを同時に行い、全ての道路エージェントのマルチモーダルな軌道を得る。計画の合理性と安全性を確保するために、衝突を考慮した再スコアモジュールを組み込んだ階層的な計画選択戦略を適用し、マルチモーダル軌道提案から最終的な計画軌道を選択する。

以上のような効果的な設計により、SparseDriveは、図1cに示すように、エンドツーエンドの自律走行の大きな可能性を解き放つ。ベルとホイッスルなしで、我々の基本モデルであるSparseDrive-Bは、平均L2エラーを19.4%(0.58m対0.72m)、衝突率を71.4%(0.06%対0.21%)大幅に削減する。従来のSOTA(最先端)手法であるUniAD[15]と比較して、我々の小さなモデルであるSparseDrive-Sは、学習で7.2倍(20時間対144時間)、推論で5.0倍(9.0FPS対1.8FPS)高速に動作し、全てのタスクで優れた性能を達成している。

我々の研究の主な貢献は以下のように要約される:

- 我々は、エンドツーエンドの自律走行のためのスパースシーン表現を探求し、スパースインスタンス表現で複数のタスクを統合するSparse-CentricパラダイムSparseDriveを提案する。

- 我々は、モーション予測とプランニングの間に共有される大きな類似性を修正し、それに対応してモーションプランナの並列設計を導く。さらに、衝突を考慮した再スコアモジュールを組み込んだ階層的な計画選択戦略を提案し、計画性能を向上させる。

- 難易度の高いnuScenes[1]ベンチマークにおいて、SparseDriveは、全てのメトリクス、特にセーフティクリティカルなメトリクス衝突率において、従来のSOTA手法を凌駕し、かつ、より高い学習効率と推論効率を維持している。

2 関連研究

2.1 マルチビュー3D検出

マルチビュー3D検出は、自律走行システムの安全性のための前提条件である。LSS[42]は、深度推定を利用して、画像特徴を3次元空間に持ち上げ、特徴をBEV平面にスプラットする。続いて、3D検出の分野にリフトスプラット操作を適用し、精度[18, 16, 25, 24]と効率[37, 17]を大幅に改善した。いくつかの作品[26, 48, 21, 5]は、BEVクエリのセットを事前に定義し、特徴サンプリングのために透視図に投影している。別の研究では、密なBEV特徴の依存性を除去している。PETRシリーズ[35, 36, 47]は、ビュー変換を暗黙的に学習するために、3次元位置エンコーディングとグローバルアテンションを導入している。Sparse4Dシリーズ[31, 32, 33]は、3D空間に明示的なアンカーを設定し、画像ビューに投影することで、局所的な特徴を集約し、繰り返しアンカーを改良する。

2.2 エンドツーエンドの追跡

ほとんどの多オブジェクト追跡(MOT)手法は、データ関連付けのような後処理に依存する、検出による追跡方式を採用している。このようなパイプラインは、ニューラルネットワークの能力を十分に活用することができない。[2]のオブジェクトクエリに触発され、いくつかの作品[52, 55, 50, 41, 46, 54]は、ストリーミング方式で追跡インスタンスをモデル化するためにトラッククエリを導入している。MOTR[52]はトラックレットを考慮したラベル割り当てを提案しており、トラッククエリは同じターゲットを継続的に検出することを強制し、検出と関連付けの間の競合に悩まされる[55, 50]。Sparse4Dv3は、時間的に伝播したインスタンスがすでに同一性の一貫性を持っていることを実証し、単純なID割り当て処理でSOTAトラッキング性能を達成する。

2.3 オンラインマッピング

オンラインマッピングは、HDマップ構築における高コストと膨大な人的労力から、HDマップの代替案として提案されている。HDMapNet[23]は、BEVセマンティックセグメンテーションを後処理でグループ化し、ベクトル化されたマップインスタンスを得る。VectorMapNet[34]は、オンライン地図構築のために2段階の自己回帰変換器を利用する。MapTR[29]はマップ要素を等価な順列の点集合としてモデル化し、マップ要素の定義の曖昧さを回避する。BeMapNetはマップ要素の詳細を記述するために区分的ベジェ曲線を採用する。StreamMapNet[51]は、時間モデリングのためのBEVフュージョンとクエリ伝搬を導入している。

2.4 エンドツーエンドの動き予測

従来のパイプラインにおけるカスケードエラーを回避するために、エンドツーエンドの動き予測を提案する。FaF[40]は、現在と将来のバウンディングボックスを予測するために、単一の畳み込みネットワークを採用している。IntentNet[3]はさらに一歩進んで、高レベルの動作と長期的な軌道の両方を推論する。PnPNet[28]は、運動予測のために軌跡レベルの特徴を集約するオンライントラッキングモジュールを導入している。ViP3D[10]は、画像とHDマップを入力として、追跡と予測を行うためにエージェントクエリを採用している。PIP[19]は、人間が注釈を付けたHDマップを局所ベクトル化マップに置き換える。

2.5 エンド・ツー・エンドの計画

エンドツーエンドプランニングの研究は、前世紀から継続されている[43]。初期の研究[6, 7, 44]では、知覚や運動予測のような中間タスクは省略されているが、これは解釈可能性に欠け、最適化が困難である。いくつかの研究[14, 4, 45, 8]は、解釈可能性を高めるために、知覚や予測結果から明示的なコストマップを構築するが、最小限のコストで最良の軌道を選択するために、手作業で作成したルールに依存している。最近、UniAD[15]は、様々なタスクをゴール指向モデルに統合する統一的なクエリ設計を提案し、知覚、予測、計画において顕著な性能を達成している。VAD[20]は、シーン学習と計画制約のためにベクトル化表現を採用している。GraphAD[56]は、交通シーンにおける複雑な相互作用のためにグラフモデルを利用する。FusionAD[49]はエンドツーエンドの駆動をマルチセンサー入力に拡張する。しかし、これまでの手法は主にシーン学習に焦点を当てており、予測と計画には、これら2つのタスクの類似性を十分に考慮することなく、素直な設計を採用しているため、性能が大きく制限されている。

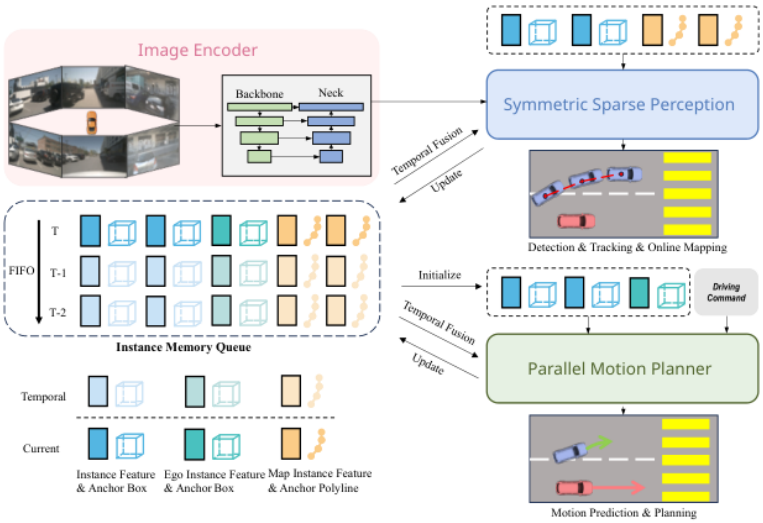


図2:SparseDriveの概要。SparseDriveは、まずマルチビュー画像を特徴マップにエンコードし、次に対称的なスパース知覚によってスパースシーン表現を学習し、最後に並列に動き予測とプランニングを行う。時間的モデリングのために、インスタンスメモリキューが考案される。

3 Method

3.1 Overview

SparseDriveの全体的なフレームワークを図2に示す。具体的には、SparseDriveは、画像エンコーダ、対称スパース知覚、並列モーションプランナの3つの部分から構成される。マルチビュー画像が与えられると、バックボーンネットワークとネックを含む画像エンコーダは、まず画像をマルチビューマルチスケール特徴マップ$I = I_s \in_R^{N\times C\times Hs\times Ws} | 1 \leqq s \leqq S$にエンコードする。対称スパース知覚モジュールでは、特徴マップIは2つのインスタンスグループに集約され、走行シーンのスパース表現を学習する。これらの2つのインスタンス群は、それぞれ周囲のエージェントとマップ要素を表し、初期化されたエゴインスタンスと対話するために、並列モーションプランナに供給される。モーションプランナは、周囲のエージェントと自車両のマルチモーダルな軌道を同時に予測し、階層的な計画選択戦略により、最終的な計画結果として安全な軌道を選択する。

3.2 対称スパース知覚

図3に示すように、スパース知覚モジュールのモデル構造は、構造的な対称性を示し、検出、追跡、オンラインマッピングを一体化している。

スパース検出。ここで、N_dはアンカーの数、Cは特徴チャネルの次元である。ここで、N_dはアンカーの数、Cは特徴チャネルの次元である。各アンカーボックスは、位置、寸法、ヨー角、速度でフォーマットされています:

$$\{x, y, z, \ln w, \ln h, \ln l, \sin yaw, \cos yaw, vx, vy, vz\}.$$

スパース検出ブランチは、単一の非時間デコーダと$N_{dec} - 1$個の時間デコーダを含むN_{dec}個のデコーダから構成される。各デコーダは特徴マップI、インスタンス特徴F_d、アンカーボックスB_dを入力とし、更新されたインスタンス特徴と洗練されたアンカーボックスを出力する。非時間デコーダはランダムに初期化されたインスタンスを入力とし、時間デコーダの入力は現在のフレームと過去のフレームの両方から来る。具体的には、非時間デコーダは、変形可能な集約、フィードフォワードネットワーク(FFN)、洗練と分類のための出力層の3つのサブモジュールを含む。変形可能な集約モジュールは、アンカーボックスB_dの周りに固定または学習可能なキーポイントを生成し、特徴サンプリングのために特徴マップIに投影する。

インスタンス特徴量F_dは、サンプリングされた特徴量との和によって更新され、分類スコアと出力層のアンカーボックスのオフセットを予測する役割を担う。時間デコーダは、最後のフレームからの時間インスタンスと現在のインスタンスの間の時間的交差注意と、現在のインスタンス間の自己注意の2つの追加のマルチヘッド注意層を持つ。マルチヘッド注意層では、アンカーボックスは高次元アンカー埋め込み$E_d \in R^{Nd \times C}$に変換され、位置符号化の役割を果たす。

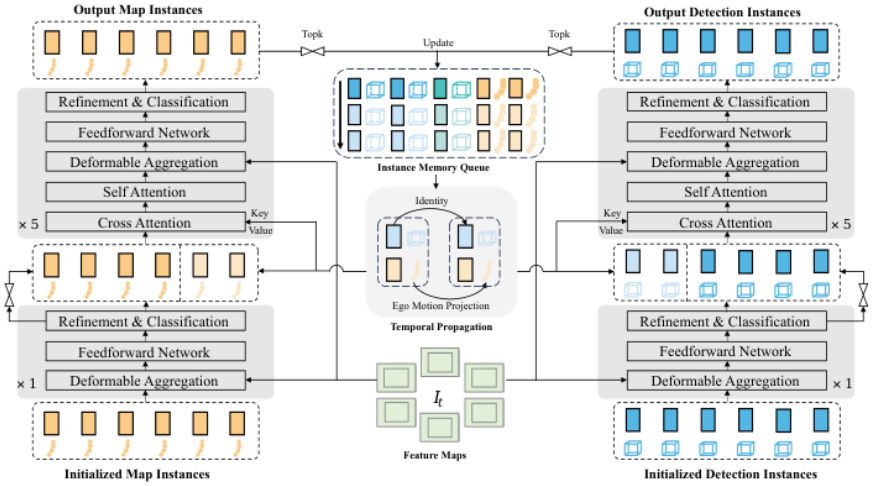


図3:対称的なスパース知覚のモデルアーキテクチャ。対称的な構造で検出、追跡、オンラインマッピングを統合している。

スパースオンラインマッピング。オンラインマッピングブランチは、インスタンスの定義が異なる以外は、検出ブランチと同じモデル構造を共有する。静的写像要素の場合、アンカーはN_p個の点を持つポリラインとして定式化される:

$$\left\{x_0, y_0, x_1, y_1, ..., x_{N_p-1}, y_{N_p-1}\right\}.$$

そして、すべての写像要素は写像インスタンス特徴$F_m \in R^{Nm \times C}$とアンカーポリライン$L_m \in R^{Nm \times Np \times 2}$で表現できる。

スパーストラッキング。トラッキングのために、Sparse4Dv3[33]のID割り当てプロセスに従う:インスタンスの検出信頼度が閾値T_{thresh}を超えると、ターゲットにロックされ、IDが割り当てられるが、これは時間伝播の間、変化しない。このトラッキング戦略はトラッキング制約を必要としないため、スパース知覚モジュールのためのエレガントでシンプルな対称設計となる。

3.3 並列モーションプランナー

図4に示すように、並列モーションプランナは、エゴインスタンス初期化、空間-時間相互作用、階層的計画選択の3つの部分から構成される。

自我インスタンスの初期化。周囲のエージェントと同様に、自車両は自車両インスタンス特徴$F_e \in R^{1 \times C}$と自車両アンカーボックス$B_e \in R^{1 \times 11}$で表現される。エゴ特徴量は、これまでの手法では一般的にランダムに初期化されていたが、我々は、エゴ特徴量は、動き予測と同様に、計画のために豊富な意味的・幾何学的情報も必要とすると主張する。しかし、周囲のエージェントのインスタンス特徴は画像特徴マップIから集約されるため、エゴ車両はカメラのブラインドエリアにあるため、実行不可能である。そこで、フロントカメラの最小特徴マップを用いて、エゴインスタンス特徴量を初期化する:

$$F_e = \text{AveragePool}(I_{front,S}) \tag{1}$$

その際、2つの利点がある。最小の特徴マップはすでに走行シーンの意味的文脈を符号化しており、密な特徴マップは、

疎な知覚では検出できないブラックリスト障害物がある場合に、疎なシーン表現の補完として機能する。

エゴアンカーB_eの場合、エゴ車両のこれらの情報を認識しているので、位置、寸法、ヨー角は自然に設定できる。速度については、[27]に示されているように、グランドトゥルースの速度から直接初期化すると、エゴの状態漏れにつながる。そこで、速度、加速度、角速度、操舵角など、現在の自我の状態ES_Tをデコードする補助タスクを追加する。各フレームにおいて、最終フレームからの予測速度をエゴアンカー速度の初期化として使用する。

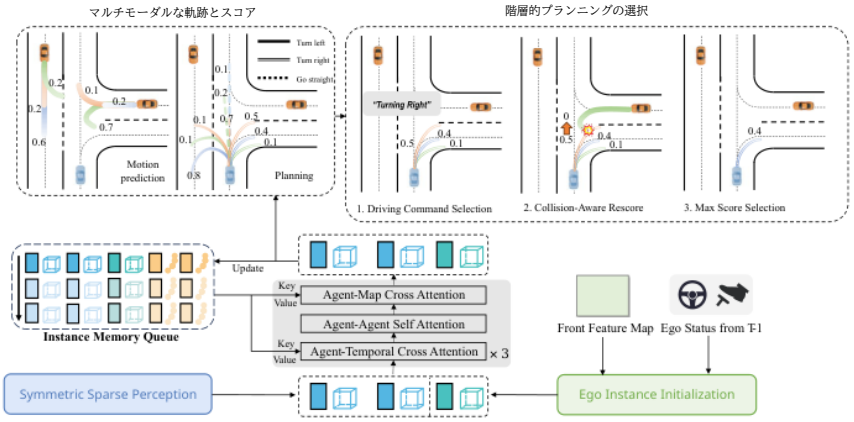


図4:並列モーションプランナのモデル構造。モーション予測とプランニングを同時に行い、安全なプランニング軌道を出力する。

時空間相互作用。すべての道路エージェント間の高レベルの相互作用を考慮するために、エゴインス we タンスを周囲のエージェントと連結してエージェントレベルのインスタンスを得る:

$$F_a = \text{Concat}(F_d, F_e), B_a = \text{Concat}(B_d, B_e) \tag{2}$$

エゴインスタンスは計画にとって重要な時間的手がかりなしに初期化されるため、時間的モデリングのために$(N_d + 1) \times H$のサイズのインスタンスメモリキューを考案する。次に、空間-時間コンテキストを集約するために、エージェント-時間クロスアテンション、エージェント-エージェントセルフアテンション、エージェント-マップクロスアテンションの3種類のインタラクションを実行する。スパース知覚モジュールの時間的交差注意では、現在のフレームのインスタンスはすべての時間的インスタンスと相互作用することに注意してください。一方、ここではエージェントと時間の交差注意のために、インスタンスレベルの相互作用を採用し、各インスタンスをそれ自身の履歴情報に集中させる。

次に、マルチモーダル軌道$\tau_m \in R^{Nd \times Km \times Tm \times 2}$、$\tau_p \in R^{Nc \times Kp \times Tp \times 2}$、スコア$s_m \in R^{Nd \times Km}$を予測する、$s_p \in R^{Ncmd \times Kp}$は周囲エージェントと自車両の両方、$K_m$と$K_p$は運動予測と計画のためのモード数、$T_m$と$T_p$は運動予測と計画のための将来のタイムスタンプ数、$N_{cmd}$は計画のための運転指令数である。一般的な慣行[15, 20]に従い、左折、右折、直進の3種類の運転コマンドを使用する。計画については、さらに自我インスタンス特徴から現在の自我状態を予測する。

階層的プランニングの選択これで、マルチモーダル計画軌道案が得られ、従うべき安全な軌道τ_p^*を1つ選択するために、階層的計画選択戦略を設計する。まず、高レベル指令cmdに対応する軌道提案の部分集合$\tau_{p,cmd} \in K_p \times T_p \times 2$を選択する。次に、安全性を確保するために、新しい衝突を考慮した再スコアモジュールを採用する。動き予測結果を用いて、各計画軌道案の衝突リスクを評価することができ、衝突確率の高い軌道については、この軌道のスコアを下げる。実際には、単純に衝突した軌跡のスコアを0に設定する。最後に、最もスコアの高い軌跡を最終的な計画出力として選択する。

3.4 エンドツーエンド学習

多段階学習。SparseDriveの学習は2つの段階に分けられる。ステージ-1では、スパースシーン表現を学習するために、対称スパース知覚モジュールをゼロから学習する。

表1: nuScenes valデータセットに対する知覚結果。SparseDriveは、エンドツーエンドの手法の中で、全ての知覚タスクで最高の性能を達成した。†: offcialチェックポイントで再現。

Method	Backbone	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
UniAD† [15]	ResNet101	0.380	0.684	0.277	0.383	0.381	0.192	0.498
SparseDrive-S	ResNet50	0.418	0.566	0.275	0.552	0.261	0.190	0.525
SparseDrive-B	ResNet101	**0.496**	**0.543**	**0.269**	**0.376**	**0.229**	**0.179**	**0.588**

(a) 3D detection results.

Method	AMOTA↑	AMOTP↓	Recall↑	IDS↓
ViP3D [10]	0.217	1.625	0.363	-
QD3DT [12]	0.242	1.518	0.399	-
MUTR3D [54]	0.294	1.498	0.427	3822
UniAD [15]	0.359	1.320	0.467	906
SparseDrive-S	0.386	1.254	0.499	886
SparseDrive-B	**0.501**	**1.085**	**0.601**	**632**

(b) 複数オブジェクトの追跡結果。

Method	AP_{ped} ↑	$AP_{divider}$ ↑	$AP_{boundry}$ ↑	mAP↑
HDMapNet [23]	14.4	21.7	33.0	23.0
VectorMapNet [34]	36.1	47.3	39.3	40.9
MapTR [29]	**56.2**	**59.8**	**60.1**	**58.7**
VAD† [20]	40.6	51.5	50.6	47.6
SparseDrive-S	49.9	**57.0**	58.4	55.1
SparseDrive-B	**53.2**	56.3	**59.1**	**56.2**

(c) Online mapping results.

ステージ2では、スパース知覚モジュールと並列モーションプランナを、モデルの重みを凍結せずに一緒に学習させ、エンドツーエンドの最適化の利点を十分に享受する。トレーニングの詳細は付録B.4に記載されている。

損失関数。損失関数は4つのタスクの損失を含み、各タスクの損失はさらに分類損失と回帰損失に分割することができる。マルチモーダルな動き予測・計画タスクには、勝者総取り戦略を採用する。計画については、エゴの状態について追加の回帰損失がある。また、知覚モジュールの学習安定性を高めるための補助タスクとして、深度推定を導入する。エンドツーエンド学習の全体的な損失関数は以下の通りである:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{map} + \mathcal{L}_{motion} + \mathcal{L}_{plan} + \mathcal{L}_{depth}. \tag{3}$$

損失関数の詳細は付録B.3に記載されている。

4 Experiments

我々の実験は、1000の複雑な運転シーンを含む、難易度の高いnuScenes[1]データセットで行われ、それぞれ約20秒間続く。各タスクの評価指標は付録Aに記述されている。このモデルには2つのバリエーションがあり、バックボーンネットワークと入力画像の解像度だけが異なる。我々の小型モデルSparseDrive-Sでは、ResNet50[11]を基幹ネットワークとして使用し、入力画像サイズは256×704である。ベースモデルであるSparseDrive-Bでは、バックボーンネットワークをResNet101に変更し、入力画像サイズを512×1408に変更した。すべての実験は8台のNVIDIA RTX 4090 24GB GPUで実施した。詳細な構成は付録Bに記載されている。

4.1 Main Results

我々は、モジュール化された方法とエンドツーエンドの方法の両方で、先行する最先端技術と比較する。エンドツーエンドの手法の中で、我々の軽量モデルSparseDrive-Sは全てのタスクで以前のSOTAを上回ったが、我々のベースモデルSparseDrive-Bは性能の境界を一歩押し上げた。各タスクの主な指標は、表中のグレーの背景で示されている。

知覚。表1aの3D検出において、SparseDriveは49.6%のmAPと58.8%のNDSを達成した。1aの3D検出において、SparseDriveは49.6%のmAPと58.8%のNDSを達成し、UniAD[15]と比較して+11.6%のmAPと+9.0%のNDSの大幅な改善をもたらした。表1bの多オブジェクト追跡では、SparseDriveは50.1%のAMOTAを達成した。1bの多オブジェクト追跡では、SparseDriveは50.1%のAMOTAを達成し、最低のIDスイッチは632で、AMOTAの点でUniAD[15]を+14.2%上回り、IDスイッチでは30.2%の削減となり、トラッキングトラックレットの時間的一貫性を示している。表1cのオンラインマッピングでは、SparseDriveは56.2%のmAPを得た。1cのオンラインマッピングでは、SparseDriveのmAPは56.2%であり、以前のエンドツーエンド手法VAD[20]を+8.6%上回っている。

表2: nuScenes valデータセットにおける動き予測とプランニングの結果。SparseDriveは従来の手法を大きく上回る。†: offcialチェックポイントで再現。* : LiDARベースの手法。

Method	minADE(m)↓	minFDE(m)↓	MR↓	EPA↑
Cons Pos. [15]	5.80	10.27	0.347	-
Cons Vel. [15]	2.13	4.01	0.318	-
Traditional [10]	2.06	3.02	0.277	0.209
PnPNet [28]	1.15	1.95	0.226	0.222
ViP3D [10]	2.05	2.84	0.246	0.226
UniAD[15]	0.71	1.02	0.151	0.456
SparseDrive-S	0.62	0.99	0.136	0.482
SparseDrive-B	**0.60**	**0.96**	**0.132**	**0.555**

(a) 予測結果

Method	L2(m)↓				Col. Rate(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
FF* [13]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO* [22]	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3 [14]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD† [15]	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
VAD† [20]	0.41	0.70	1.05	0.72	0.03	0.19	0.43	0.21
SparseDrive-S	**0.29**	0.58	0.96	0.61	**0.01**	0.05	0.18	0.08
SparseDrive-B	**0.29**	**0.55**	**0.91**	**0.58**	**0.01**	**0.02**	**0.13**	**0.06**

(b) Planning results.

予測。表2aの動き予測では、SparseDriveが最も良い性能を達成した。2aの動き予測では、SparseDriveは0.60m minADE、0.96m minFDE、13.2% MissRate、0.555 EPAで最高の性能を達成した。UniAD[15]と比較して、SparseDriveはminADEとminFDEでそれぞれ15.5%と5.9%の誤差を削減する。

プランニング。Tab.2bのプランニングでは、すべての手法の中で SparseDriveは、L2誤差0.58m、衝突率0.06%と、全ての手法の中で顕著なプランニング性能を達成している。SparseDriveは従来のSOTA VAD[20]と比較して、L2誤差を19.4%、衝突率を71.4%低減し、本手法の有効性と安全性を実証した。

効率。表3に示すように、SparseDriveは、SparseDriveよりも効率的である。3に示すように、SparseDriveは優れた性能の他に、学習と推論の両方ではるかに高い効率を達成している。同じバックボーンネットワークで、我々のベースモデルはUniAD[15]と比較して、学習で4.8倍、推論で4.1倍高速に達成する。我々の軽量モデルは、学習と推論において7.2倍と5.0倍高速に達成できる。

表 3: 効率の比較結果。SparseDriveは学習と推論の両方で高い効率を達成している。UniADの学習時間とFPSは、それぞれ8GPUと1GPUのNVIDIA Tesla A100で測定。SparseDriveの学習時間とFPSは、それぞれ8GPUと1GPUのNVIDIA Geforce RTX 4090で測定。

Method	Training Efficiency			Inference Efficiency			
	GPU Memory (G)	Batch Size	Time (h)	GPU Memory (M)	FLOPs (G)	Params (M)	FPS
UniAD [15]	50.0	1	48 + 96	2451	1709	125.0	1.8
SparseDrive-S	15.2	6	18 + 2	1294	192	85.9	9.0
SparseDrive-B	17.6	4	26 + 4	1437	787	104.7	7.3

4.2 アブレーション研究

我々は、設計上の選択の有効性を実証するために、広範なアブレーション研究を実施した。アブレーション実験のデフォルトモデルとしてSparseDrive-Sを使用する。

モーションプランナーにおけるデザインの効果予測と計画の類似性を考慮することの重要性を強調するために、我々はいくつかの具体的な実験を考案した。4. ID-2は、予測と計画の並列設計を逐次順序に変更することで、エゴ車両が周囲のエージェントに与える影響を無視し、動き予測と衝突率の性能を悪化させる。ID-3は自我インスタンスの特徴をランダムに初期化し、自我アンカーの全パラメータを0に設定する。自我インスタンスの意味的・幾何学的情報を削除すると、L2誤差と衝突率の両方で性能劣化が生じる。ID-4はプランニングを決定論的な問題として捉え、ある軌跡を1つだけ出力するため、衝突率が最も高くなる。さらに、ID-5はインスタンスレベルのエージェント-時間交差注意を除去し、L2誤差を0.77mに深刻に劣化させる。衝突を考慮した再スコアについては、以下の段落で詳細な議論を行う。

衝突を考慮したスコア。これまでの手法[15, 56]では、知覚結果に基づいて安全性を確保するために、ポスト最適化戦略が採用されている。しかし、我々は、この戦略がエンドツーエンドのパラダイムを破り、その結果、L2エラーが深刻な劣化をもたらすと主張する。5. さらに、我々の再実装された衝突率メトリックの下では、ポスト最適化はプランニングをより安全にするのではなく、むしろより危険なものにする。一方、衝突を考慮した再スコアモジュールは、衝突率を0.12%から0.08%に低減し、L2誤差を無視できるほど増加させ、本手法の優位性を示している。

マルチモーダルプランニング。プランニングモードの数について実験を行う。表6に示すように 6に示すように、プランニングモード数が増加するにつれて、プランニング性能は6モードで飽和するまで継続的に向上し、マルチモーダルプランニングの重要性が改めて証明された。

表4:並列モーションプランナにおける設計のアブレーション。”PAL ”は動作予測・計画タスクの並列設計、”EII ”はエゴインスタンス初期化、”MTM ”は計画のための複数モード、”ATA ”はエージェント-時間交差注意、”CAR ”は衝突を考慮した再スコアを意味する。

ID	PAL	EII	MTM	ATA	CAR	Prediction			Planning L2(m)				Planning Coll.(%)			
						minADE	minFDE	MR	1s	2s	3s	Avg.	1s	2s	3s	Avg.
1	✓	✓	✓	✓	✓	0.623	**0.987**	0.136	**0.29**	**0.58**	0.96	**0.61**	**0.01**	**0.05**	**0.18**	**0.08**
2		✓	✓	✓	✓	0.641	1.008	0.138	0.30	0.58	0.95	**0.61**	0.02	0.06	0.23	0.10
3	✓		✓	✓	✓	**0.621**	0.988	**0.135**	0.31	0.60	0.98	0.63	0.03	0.07	0.21	0.11
4	✓	✓		✓	✓	0.626	1.002	0.136	0.33	0.66	1.08	0.69	0.03	0.11	0.60	0.25
5	✓	✓	✓		✓	0.634	1.003	0.138	0.40	0.74	1.16	0.77	0.02	0.13	0.32	0.16
6	✓	✓	✓	✓		0.623	**0.987**	0.136	**0.29**	**0.58**	**0.95**	**0.61**	**0.01**	0.06	0.30	0.12

表5:[15]の衝突を考慮した再スコアとポスト最適化のためのアブレーション。

Method	CAR	Post-optim.	Planning L2(m)				Planning Coll.(%)			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD[15]			0.32	0.58	0.94	0.61	0.15	0.24	0.36	0.25
UniAD[15]		✓	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
SparseDrive			**0.29**	**0.58**	**0.95**	**0.61**	**0.01**	0.06	0.30	0.12
SparseDrive	✓		**0.29**	**0.58**	0.96	**0.61**	**0.01**	**0.05**	**0.18**	**0.08**
SparseDrive		✓	0.44	0.73	1.11	0.76	0.29	0.21	0.38	0.30

表6:プランニングモードのアブレーション

Number of mode	Planning L2(m)				Planning Coll.(%)			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
1	0.33	0.66	1.08	0.69	0.03	0.11	0.60	0.25
2	0.33	0.65	1.08	0.69	0.01	0.12	0.42	0.18
3	0.30	0.59	0.97	0.62	**0.00**	0.08	0.43	0.17
6	**0.29**	**0.57**	**0.95**	**0.61**	0.01	**0.03**	**0.17**	**0.07**
9	0.33	0.63	1.04	0.66	0.01	0.09	0.36	0.15

5 結論と今後の課題

結論。本研究では、疎なシーン表現を探求し、エンドツーエンドの自律走行の領域におけるタスク設計をレビューする。エンドツーエンドのパラダイムであるSparseDriveは、顕著な性能と高い効率の両方を達成する。SparseDriveの素晴らしい性能が、エンドツーエンドの自律走行のためのタスク設計を再考し、この分野の技術進歩を促進するコミュニティを刺激することを願っている。

今後の課題である。我々の研究にはまだいくつかの限界がある。第一に、我々のエンドツーエンドモデルの性能は、例えばオンラインマッピングタスクのようなシングルタスク手法にまだ及ばない。第二に、データセットの規模がエンドツーエンドの自律走行の可能性を十分に活用できるほど大きくなく、オープンループ評価ではモデル性能を包括的に表現できない。これらの問題は今後の検討に委ねる。

References

[1] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018.

[4] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021.

[5] Shaoyu Chen, Tianheng Cheng, Xinggang Wang, Wenming Meng, Qian Zhang, and Wenyu Liu. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv preprint arXiv:2206.04584*, 2022.

[6] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4693–4700. IEEE, 2018.

[7] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9329–9338, 2019.

[8] Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, and Raquel Urtasun. Lookout: Diverse multi-future prediction and planning for self-driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16107–16116, 2021.

[9] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[10] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1992–2008, 2022.

[13] Peiyun Hu, Aaron Huang, John Dolan, David Held, and Deva Ramanan. Safe local motion planning with self-supervised freespace forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12732–12741, 2021.

[14] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022.

[15] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[16] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

[17] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022.

[18] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

[19] Bo Jiang, Shaoyu Chen, Xinggang Wang, Bencheng Liao, Tianheng Cheng, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, and Chang Huang. Perceive, interact, predict: Learning dynamic and static clues for end-to-end motion prediction. *arXiv preprint arXiv:2212.02181*, 2022.

[20] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023.

[21] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 1042–1050, 2023.

[22] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *European Conference on Computer Vision*, pages 353–369. Springer, 2022.

[23] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634. IEEE, 2022.

[24] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1486–1494, 2023.

[25] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1477–1485, 2023.

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[27] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? *arXiv preprint arXiv:2312.03031*, 2023.

[28] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020.

[29] リャオ・ベンチェン、チェン・シャオユー、ワン・シンガン、チェン・ティエンヘン、チャン・チアン、リウ・ウェンユー、チャン・ファン。Mapstr: オンラインベクトル化hdマップ構築のための構造化モデリングと学習。第11回学習表現国際会議(2022年)にて。

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[31] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

[32] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023.

[33] Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*, 2023.

[34] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023.

[35] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[36] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023.

[37] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[38] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[41] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022.

[42] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020.

[43] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[44] アディティヤ・プラカシュ、カシャップ・チッタ、アンドレアス・ガイガー。エンドツーエンドの自律走行のためのマルチモーダル融合変換器。コンピュータビジョンとパターン認識に関するIEEE/CVF会議予稿集，ページ7077-7087，2021.

[45] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 414–430. Springer, 2020.

[46] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[47] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023.

[48] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023.

[49] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*, 2023.

[50] Ei劉、王天才、李珠玲、張玉陽、張仙玉、田文平。Motrv3：エンドツーエンドの多オブジェクト追跡のためfeのリリーススーパービジョン。arXivプレプリント arXiv:2305.14298，2023.

[51] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024.

[52] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022.

[53] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023.

[54] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2022.

[55] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023.

[56] Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. *arXiv preprint arXiv:2403.19098*, 2024.

A Metrics

検出と追跡の評価は標準的な評価プロトコル[1]に従う。検出には、平均平均精度(mAP)、平均並進誤差(mATE)、スケール(mASE)、オリエンテーション(mAOE)、速度(mAVE)、属性(mAAE)、nuScenes検出スコア(NDS)を用いて、モデル性能を評価する。トラッキングには、平均多オブジェクトトラッキング精度(AMOTA)、平均多オブジェクトトラッキング精度(AMOTP)、RECALL、IDスイッチ(IDS)をメトリクスとして使用する。オンラインマッピングでは、車線分割、横断歩道、道路境界の3つのマップクラスの平均精度(AP)を計算し、全クラスの平均精度(mAP)を求める。動き予測には、[10]で提案された最小平均変位誤差(minADE)、最小最終変位誤差(minFDE)、ミス率(MR)、エンドツーエンド予測精度(EPA)などの指標を採用する。動き予測ベンチマークはUniAD[15]と連携している。

プランニングには、一般的に用いられるL2誤差と衝突率を採用し、プランニング性能を評価する。L2誤差の評価はVAD[20]と一致する。衝突率については、以前の[15, 20]の実装では2つの欠点があり、その結果、計画性能の評価が不正確になる。一方、従来のベンチマークでは、障害物のバウンディングボックスをグリッドサイズ0.5mの占有マップに変換しており、例えば、自車両が1つの占有マップピクセルより小さい障害物に接近するなど、特定のケースで誤った衝突が発生している[53]。(2) 自車両の方位は考慮されず、変更されないと仮定される[27]。計画性能を正確に評価するために、軌跡点を通るヨー角を推定することで自我方位の変化を考慮し、自我車両と障害物のバウンディングボックスの重なりを調べることで衝突の有無を評価する。我々のベンチマークにおける計画結果を公式チェックポイント[15, 20]で再現し、公平に比較する。

B 実装の詳細

B.1 Perception

スパース知覚モジュールでは、デコーダ層の数N_{dec}を6とし、非時間デコーダ1個、時間デコーダ5個とする。アンカーボックスB_dとアンカーポリラインL_mの位置は、学習セットに対するK-Meansクラスタリングによって得られ、アンカーボックスの他のパラメータは{1, 1, 1, 0, 1, 0, 0}で初期化される。各マップ要素は20点で表現される。アンカーボックスの数N_dとポリラインの数N_mはそれぞれ900と100に設定され、検出とオンラインマッピングの時間インスタンスの数は600と33である。トラッキング閾値T_{thresh}は0.2に設定される。検出のために、知覚範囲は半径55mの円である。オンラインマッピングの場合、知覚範囲は縦横60m×30mである。マルチヘッドアテンションには、GPUメモリを節約するためにFlash Attention[9]を採用する。

B.2 Motion Planner

インスタンスメモリキューのストアフレーム数Hは3である。動き予測のためのモード数K_mとプランニングのためのモード数K_pは共に6とする。動き予測のための未来のタイムスタンプ数T_mとプランニングのための将来のタイムスタンプ数T_pはそれぞれ12と6とする。モーションプランナにおける空間-時間相互作用の後、多層パーセプトロン(MLP)を用いて、自我特徴F_eを持つ現在のフレームの自我状態をデコードする:

$$ES_T = MLP(F_e) \tag{4}$$

マルチモーダルな軌跡とスコアについては、K-Meansクラスタリングを用いて事前意図点を求め、それらを動きモードクエリM $Q_m \in R^{Km \times C}$ と正弦波位置エンコーディングPE(-)を用いた計画モードクエリM $Q_p \in R^{Ncmd \times Kp \times C}$ に変換し、エージェントインスタンス特徴によるモードクエリを追加し、MLPで軌跡とスコアをデコードする:

$$\tau_m = MLP(F_d + MQ_m), \tag{5}$$
$$s_m = MLP(F_d + MQ_m), \tag{6}$$
$$\tau_p = MLP(F_e + MQ_p), \tag{7}$$
$$s_p = MLP(F_e + MQ_p) \tag{8}$$

衝突を考慮した再スコアモジュールでは、動き予測において最も信頼性の高い2つの軌道を利用し、エゴ車両が周囲の障害物に衝突するかどうかを判断する。

B.3 損失関数

知覚については、ハンガリーアルゴリズムが各グランドトゥルースを1つの予測値でマッチングさせるために採用される。検出損失は、分類のためのFocal損失[30]

$$L_{det} = \lambda_{det\_cls} L_{det\_cls} + \lambda_{det\_reg} L_{det\_reg}. \tag{9}$$

とボックス回帰のためのL1損失の線形結合である：ID割り当て処理には追跡制約がないため、追跡損失はない。オンラインマッピングの損失は検出の損失と同様である。

$$L_{map} = \lambda_{map\_cls} L_{map\_cls} + \lambda_{map\_reg} L_{map\_reg}. \tag{10}$$

深度推定には、回帰にL1損失を用いる：

$$L_{depth} = \lambda_{depth} L_{depth}. \tag{11}$$

λ_{det_cls}=2、λ_{det_reg}=0.25、λ_{map_cls}=1、λ_{map_reg}=10、λ_{depth}=0.2。

動き予測とプランニングのために、マルチモデル出力とグランドトゥルース軌道の間の平均変位誤差(ADE)を計算し、ADEが最も低い軌道を正サンプル、それ以外を負サンプルとする。計画については、自我の状態も追加で予測される。また、分類にはFocal lossを、回帰にはL1 lossを用いる：

$$\begin{aligned} L_{motion\_planning} &= \lambda_{motion\_cls} L_{motion\_cls} + \lambda_{motion\_reg} L_{motion\_reg} \\ &+ \lambda_{plan\_cls} L_{plan\_cls} + \lambda_{plan\_reg} L_{plan\_reg} + \lambda_{plan\_status} L_{plan\_status}, \end{aligned} \tag{12}$$

where $\lambda_{motion\_cls} = 0.2$, $\lambda_{motion\_reg} = 0.2$, $\lambda_{plan\_cls} = 0.5$, $\lambda_{plan\_reg} = 1.0$, $\lambda_{plan\_status} = 1.0$.

B.4 Training Details

モデル学習にはAdamWオプティマイザ[39]とコサインアニーリング[38]スケジューラを用いる。学習ハイパーパラメータをTab. 7.

表7:学習ハイパーパラメータ。

Model	Training stage	Batch Size	Epochs	Lr	Backbone lr scale	Weight decay
SparseDrive-S	stage-1	8	100	4×10^{-4}	0.5	1×10^{-3}
SparseDrive-S	stage-2	6	10	3×10^{-4}	0.1	1×10^{-3}
SparseDrive-B	stage-1	4	80	3×10^{-4}	0.1	1×10^{-3}
SparseDrive-B	stage-2	4	10	3×10^{-4}	0.1	1×10^{-3}

C Visualization

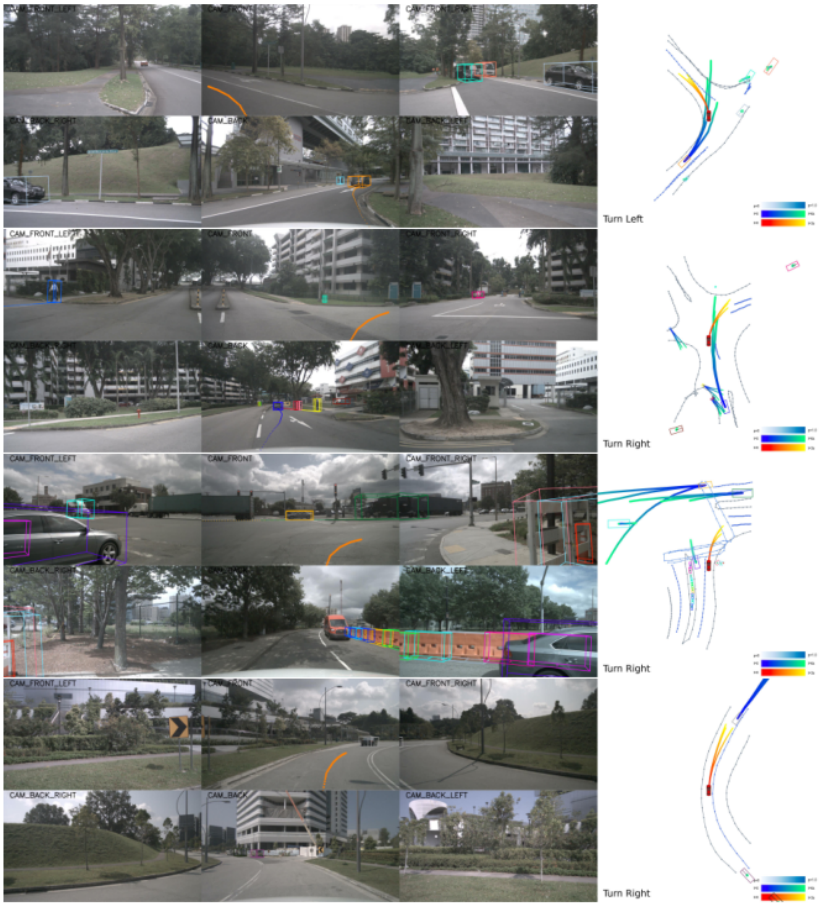


図5:可視化結果。SparseDriveは交差点で異なる旋回モードを学習する。

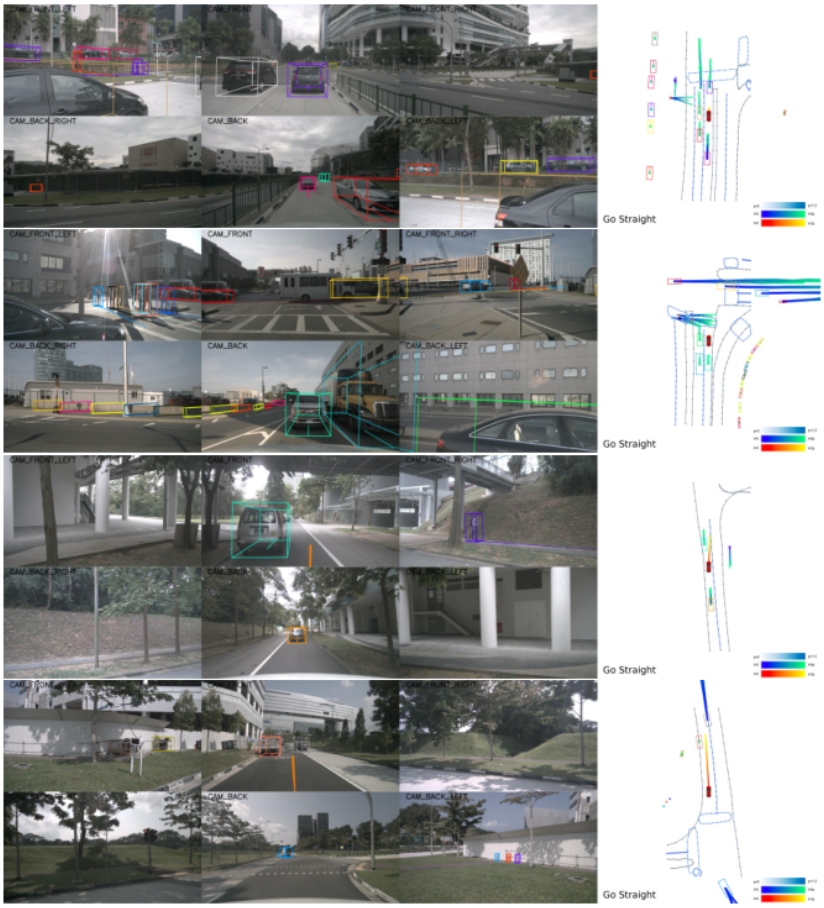


図6:可視化結果。SparseDriveは、移動するエージェントに譲り渡す、または障害物との衝突を回避することを学習する。