ParkingE2E: カメラベースのエンドツーエンドのパーキングネットワーク、画像から計画へ

Changze Li, Ziheng Ji, Zhe Chen, Tong Qin*, and Ming Yang

概要- 自律駐車は、インテリジェント運転分野において重要なタスクである。従来の駐車アルゴリズムは、通常、ルールベースのスキームを用いて実装されている。しかし、これらの方法は、アルゴリズムの設計が複雑であるため、複雑な駐車シナリオではあまり有効ではない。一方、ニューラルネットワークベースの手法は、ルールベースの手法よりも直感的で汎用性が高い傾向にある。エキスパートによる駐車軌跡のデータを大量に収集し、学習ベースの手法で人間戦略をエミュレートすることで、駐車タスクに効果的に対処することができる。本論文では、模倣学習を採用し、人間の運転軌跡を模倣することで、RGB画像から経路計画へのエンドツーエンドのプランニングを行う。提案するエンドツーエンドのアプローチは、画像とターゲット特徴を融合するためにターゲットクエリエンコーダを利用し、将来のウェイポイントを自己回帰的に予測するために変換器ベースのデコーダを利用する。実世界のシナリオで広範な実験を行い、その結果、提案手法は4つの異なる実世界のガレージにおいて、平均87.8%の駐車成功率を達成した。実車実験により、本論文で提案した手法の実現可能性と有効性がさらに検証された。コードはhttps://github.com/qintonguav/ParkingE2Eにある。

I. Introduction

インテリジェント運転には、都市運転、高速道路運転、駐車操作の3つの主要タスクが含まれる。自動バレーパーキング(AVP)と自動駐車支援(APA)システムは、インテリジェント運転における重要な駐車タスクであり、駐車の安全性と利便性を大幅に改善する。しかし、主流の駐車方法[1]はルールベースであることが多く、駐車プロセス全体を環境認識、マッピング、スロット検出、ローカライゼーション、経路計画などの複数のステージに分解する必要がある。このような複雑なモデル・アーキテクチャは複雑な性質を持っているため、狭い駐車場や複雑なシナリオで困難に遭遇しやすい。

エンドツーエンド(E2E)自律走行アルゴリズム[3]-[7]は、知覚、予測、計画の各コンポーネントを統合して、共同最適化のための統一的なニューラルネットワークに統合することで、モジュール間の累積誤差を軽減する。エンドツーエンドのアルゴリズムを駐車場シナリオに適用することで、駐車場システムの手作業で設計された機能やルールへの依存度を下げ、包括的で全体的でユーザーフレンドリーなソリューションを提供することができる。

エンドツーエンドの自律走行は大きな利点を示しているが、ほとんどの研究は、アルゴリズムの実世界での有効性を検証することなく、シミュレーション[8]に集中している。

すべての著者は、上海交通大学未来技術グローバル研究所(中国・上海)に所属している。{changze, jiziheng, Zhe Chen, qintong, mingyang}@sjtu.edu.cn.* は対応する著者である。

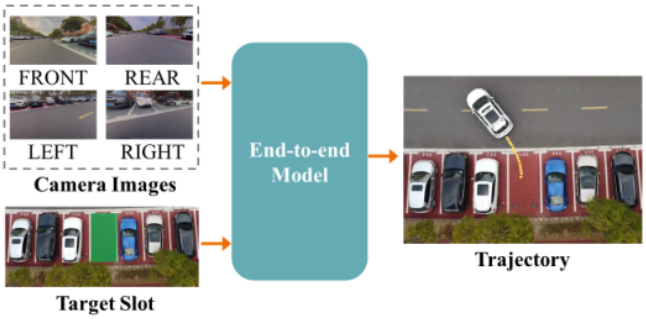

図1:全体的なワークフローの説明図。我々のモデルは、サラウンドビューカメラ画像とターゲットスロットを入力とし、予測された軌跡ウェイポイントを出力し、後にコントローラによって実行される。補足動画は https://youtu.be/urOEHJH1TBQ で入手可能である。

都市環境の複雑さや高速道路走行の危険性とは対照的に、駐車場のシナリオは、低速、閉鎖空間、高い制御性が特徴である。これらの特徴は、エンド・ツー・エンドの自律走行機能を自動車に漸進的に導入するための実現可能な道筋を提供する。エンドツーエンドの駐車ニューラルネットワークを開発し、実世界の駐車状況におけるアルゴリズムの実現可能性を検証する。

本研究では、模倣学習ベースのエンドツーエンドの駐車アルゴリズムを提示することで、我々の以前の研究E2E-Carla [2]を拡張し、実環境での展開と評価に成功した。このアルゴリズムは、搭載カメラで撮影されたサラウンドビュー画像を取り込み、将来の軌跡結果を予測し、予測されたウェイポイントに基づいて制御を実行する。ユーザーが駐車場を指定すると、エンドツーエンドの駐車場ネットワークがコントローラと連携し、完全に駐車されるまで自動的に駐車場に車両を操作する。本論文の貢献は以下のようにまとめられる。

- 駐車場タスクを実行するために、エンドツーエンドのネットワークを設計した。ネットワークは、周囲のビュー画像を鳥瞰図(BEV)表現に変換し、ターゲット特徴を画像特徴へのクエリに採用することで、ターゲット駐車場の特徴と融合させる。軌跡点の逐次的な性質のため、軌跡点を生成するために、変換デコーダに基づく自己回帰アプローチを利用する。
- 我々は、テストのために実車上にエンドツーエンドモデルを配備し、様々な実世界のシナリオにわたって駐車のためのネットワークモデルの実現可能性と一般性を検証し、エンドツーエンドのネットワーク展開のための効果的なソリューションを提供する。

II. LITERATURE REVIEW

A. BEV知覚

BEV表現は、遠近法表現と比較して、少なくとも2つの利点がある。まず、物理的な解釈が可能であるため、異なるモダリティからの入力を容易に統合することができる。第二に、BEVビューは遠近法の歪みの問題を回避し、それによってプランニングのような下流タスクの複雑さを軽減する。近年、BEV表現は知覚システムにおいて広く採用されている。特徴抽出モジュールとタスクヘッドモジュールから構成される従来の深層学習ベースの知覚アルゴリズムとは異なり、BEV知覚は、これら2つのモジュールと並行して、追加の視点変換モジュールを組み込んでいる。この変換モジュールは、センサービューと鳥瞰図(BEV)の間の変換を容易にする。

LSS[23]は、検出とセグメンテーションにBEV知覚を利用する。本手法は、特徴マップの各画素における奥行き分布を推定し、BEV平面に投影することで、BEV特徴量を取得する。DETR3D [26]はDETR [25]の基本パラダイムに従い、3Dオブジェクト検出のためにスパースクエリを採用する。PETR[27]は、3次元位置埋め込みを追加し、2次元特徴量に3次元位置情報を提供し、ニューラルネットワークが暗黙的に奥行きを学習することを目指す。BEVFormer[28]は知覚にBEVクエリを採用し、空間的交差注意と時間的自己注意メカニズムを組み込んで性能を向上させている。BEVDepth[29]はLSSをベースに、学習時にLiDARポイントを奥行き監視に利用することで、奥行き推定品質を向上させ、BEV知覚性能を向上させる。BEVFusion [30]は、カメラとLiDARデータの両方からBEV特徴を抽出し、BEV空間で融合する。

B. エンドツーエンドの自律走行

従来のモジュールベースの自律走行ソリューションとは対照的に、エンドツーエンドのパラダイム[9, 10]は、蓄積されたエラーを軽減し、モジュール間の情報損失を防ぎ、冗長な計算を最小化することができる。その結果、自律走行タスクの分野で人気のある著名な研究テーマとして浮上してきた。

エンド・ツー・エンドの運転に関する研究は、当初は自律的な都市運転タスクに焦点を当てていた。模倣学習に基づくエンドツーエンド手法である ChauffeurNet [11]は、専門家のデータから効果的な運転戦略を学習した。多くの手法が、センサーからBEV特徴を抽出し、GRU(Gate Recurrent Unit)デコーダーを利用して、Transfuser [3, 12]、Interfuser [13]、NEAT [14]などの自己回帰的にウェイポイントを予測するエンコーダ・デコーダのフレームワークを採用している。また、CIL[15]とCILRS[16]は、独立したPIDコントローラを用いずに、フロントビュー画像、電流測定値、ナビゲーションコマンドを制御信号に直接マッピングするニューラルネットワークを開発した。MP3[17]とUniAD[7]はモジュール設計を提案しているが、エンドツーエンドで全てのコンポーネントを共同最適化する。

近年、駐車シナリオのためのエンドツーエンドネットワークが開発されている。Rathourら[18]は、画像から操舵角と歯車を予測する2段階学習フレームワークを提案した。第一段階では、ネットワークは一連の操舵角の初期推定値を予測する。第2段階では、LSTM(Long Short-Term Memory)ネットワークを使用して、最適なステアリング角度とギアを推定する。Liら[19]は、ステアリングの角度と速度を自動的に制御するために、後景画像に対してCNN(畳み込みニューラルネットワーク)を学習させた。ParkPredict[20]は、CNN-LSTMアーキテクチャに基づく駐車場スロットとウェイポイント予測ネットワークを提案した。以下の研究では、ParkPredict+ [21]が、意図、画像、過去の軌跡に基づいて将来の車両軌跡を予測する変換器とCNNベースのモデルを設計した。

既存のエンドツーエンドの自律走行手法は、多くの場合、多大な計算資源を必要とし、トレーニングの課題を提起し、実車展開の困難に直面している。一方、ParkPredictに代表される駐車アプローチは、主に航空画像からの予測に重点を置いており、我々のタスクとは異なる。本手法は、RGB画像とターゲットスロットから抽出されたBEV特徴から将来のウェイポイントを予測するために、自己回帰変換デコーダを利用するエンドツーエンドの駐車計画ネットワークを提案する。

III. METHODOLOGY

A. 前置き：問題の定義

エンドツーエンドのニューラルネットワークN_θを用いて、エキスパートの軌跡を模倣して学習し、データセットを定義する：

$$\mathcal{D} = \{(I^k_{i,j}, P_{i,j}, S_i)\}, \tag{1}$$

ここで、軌跡インデックスi∈[1, M]、軌跡点インデックスj∈[1, N_i]、カメラインデックスk∈[1, R]、RGB画像I、軌跡点P、ターゲットスロットSである：

$$\mathcal{T}_{i,j} = \{P_{i,\min(j+b,N_i)}\}_{b=1,2,\ldots,Q}, \tag{2}$$

and

$$\mathcal{D}' = \{(I^k_{i,j}, \mathcal{T}_{i,j}, S_i)\}, \tag{3}$$

ここで、Qは予測された軌跡点の長さを表し、RはRGBカメラの数を表す。

エンドツーエンドネットワークの最適化目標は以下の通りである：

$$\theta' = \arg\min_\theta \mathbb{E}_{(I,\mathcal{T},S)\sim\mathcal{D}'}[\mathcal{L}(\mathcal{T}, \mathcal{N}_\theta(I,S))], \tag{4}$$

ここで、Lは損失関数を表す。

B. カメラベースのエンドツーエンドニューラルプランナー

1）概要 RGB画像とターゲットスロットを入力とするエンドツーエンドのニューラルプランナを開発した。提案するニューラルネットワークは、入力エンコーダと自己回帰軌跡デコーダの2つの主要部分から構成される。RGB画像とターゲットスロットを入力として、RGB画像はBEV特徴に変換される。次に、ニューラルネットワークはBEV特徴をターゲットスロットと融合し、変換デコーダを用いて自己回帰的に次の軌跡点を生成する。

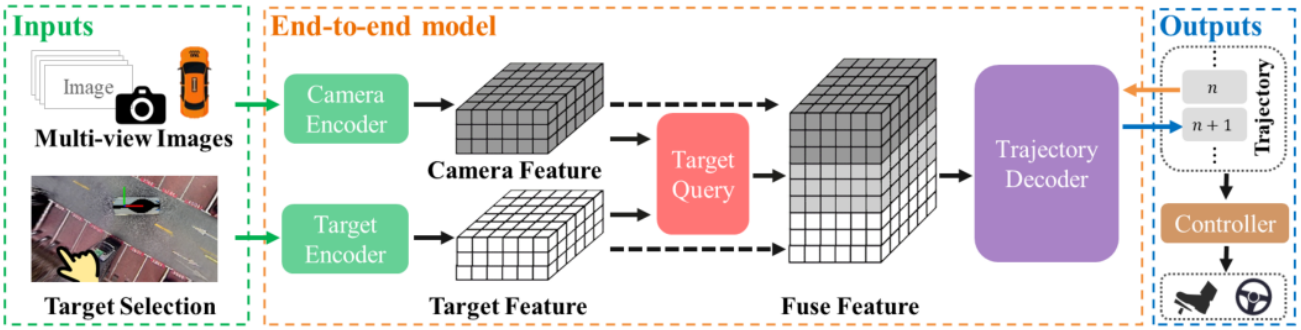

図2:本手法の概要。マルチビューRGB画像が処理され、画像特徴がBEV表現に変換される。ターゲットスロットは、BEV ターゲット特徴を生成するために使用される。ターゲットクエリを用いて、ターゲット特徴量と画像BEV特徴量を融合する。次に、自己回帰変換デコーダを用いて、予測された軌跡点を一つずつ求める。

2) エンコーダ: 入力をBEVビューにエンコードする。BEV表現は、車両の周囲環境をトップダウンで表示し、自車両が駐車スロット、障害物、マーキングを検出することを可能にする。同時に、BEVビューは様々な運転視点にわたって一貫した視点表現を提供し、それによって軌道予測の複雑さを単純化する。
カメラエンコーダ BEV生成パイプラインの最初に、まずEfficientNet [22]を利用して、RGB入力から画像特徴F_{img}∈_{R}^{C×Himg×Wimg}を抽出する。LSS[23]にヒントを得て、画像特徴の深度分布d_{dep}∈_{R}^{D×Himg×Wimg}
を学習し、各画素を3次元空間に持ち上げる。次に、予測された奥行き分布d_{dep}と画像特徴F_{img}を掛け合わせ、奥行き情報を持つ画像特徴を得る。カメラのエクストリンシックとイントリンシックにより、画像特徴をBEVボクセルグリッドに投影し、カメラ特徴F_{cam}∈_{R}^{C×Hcam×Wcam}
を生成する。x方向のBEV特徴量の範囲を[-R_x, R_x]m、ここでmはメートルを表し、y方向の範囲を[-R_y, R_y]mと表す。

ターゲットエンコーダ ターゲットスロットをカメラ特徴量F_{cam}に合わせるために、指定された駐車スロット位置に基づいて、ターゲットエンコーダの入力としてBEV空間のターゲットヒートマップを生成する。その後、ディープCNNニューラルネットワークを用いてターゲットスロット特徴量F_{target}を抽出し、F_{cam}と同じ次元を得る。学習中、目標駐車場のスロットは、人間の運転軌跡の終点によって決定される。

ターゲットクエリ BEV空間において、カメラ特徴量F_{cam}とターゲット符号化特徴量F_{target}を整列させ、ターゲット特徴量を用いてクロスアテンションメカニズムによりカメラ特徴量を問い合わせることで、2つのモダリティを効果的に融合することができる。位置エンコーディングは、特定のBEV位置で特徴を関連付ける際に、カメラ特徴とターゲット特徴の間の空間的対応が維持されることを保証する。F_{target}をクエリ、カメラ特徴量F_{cam}をキー、値として利用し、注目機構を用いることで、融合特徴量F_{fuse}を得ることができる。

3) デコーダ: 多くのエンドツーエンド計画研究[12]-[14]では、GRUデコーダを用いて、以下の点から次の点を予測している。

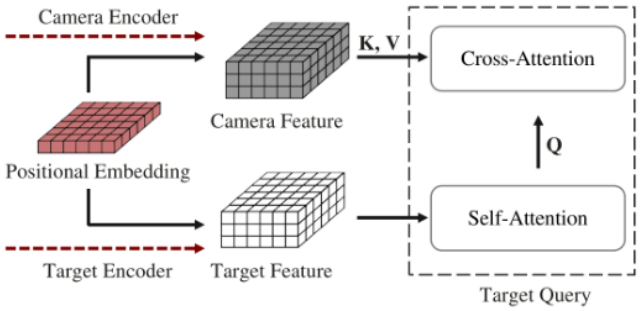

図3:ターゲットクエリのアーキテクチャは、2種類の特徴間の空間的関係を確立するために、ターゲット特徴とカメラ特徴に同じ位置エンコーディングを追加することを示している。

高次元特徴ベクトルを自己回帰的に学習する。しかし、特徴量の高次元ベクトルはグローバルな受容野を持たない。Pix2seq[24]からヒントを得て、我々は変換デコーダを用いたシーケンス予測問題として軌道計画にアプローチする。これは、軌跡点の自己回帰的なステップバイステップの予測を含む。本アプローチは、低次元の軌跡点と高次元の画像特徴を効果的に組み合わせる。
Trajectory Serialization Trajectory Serializationは軌跡点を離散トークンとして表現する。軌跡点を直列化することで、位置回帰をトークン予測に変換することができる。その後、変換デコーダを利用して、自車両の座標系における軌跡点(P_{ij}^{x}, P_{ij}^{y})を予測するために、以下の直列化手法を利用する:

$$\mathrm{Ser}(P_{i,j}^{x}) = \lfloor \frac{P_{i,j}^{x} + R_x}{2R_x} \rfloor \times N_t, \tag{5}$$

and

$$\mathrm{Ser}(P_{i,j}^{y}) = \lfloor \frac{P_{i,j}^{y} + R_y}{2R_y} \rfloor \times N_t, \tag{6}$$

ここで、N_tはシーケンス中のトークンによって符号化できる最大値を表し、軌跡点を直列化する記号をSer(-)と表す。R_xとR_yはそれぞれx方向とy方向の予測範囲の最大値を表す。

直列化後、i番目の軌跡は以下のように表現できる：

$$[\mathrm{BOS}, \mathrm{Ser}(P^x_{i,1}), \mathrm{Ser}(P^y_{i,1}), ..., \mathrm{Ser}(P^x_{i,N_i}), \mathrm{Ser}(P^y_{i,N_i}), \mathrm{EOS}], \tag{7}$$

ここで、BOSは開始フラグ、EOSは終了フラグを表す。

軌跡デコーダ BEV特徴量をキーと値とし、直列化シーケンスをクエリとして利用し、自己回帰的に変換デコーダを用いて軌跡点を生成する。学習時には、シーケンス点に位置埋め込みを追加し、未知の情報をマスキングすることで並列化を実現する。推論プロセスにおいて、BOSトークンが与えられると、変換デコーダは以下の点を順番に予測する。次に、EOSに遭遇するか、指定された数の予測点に到達するまで、このプロセスを繰り返す次のステップのために、予測点をシーケンスに追加する。

C. 横方向および縦方向の制御

制御過程において、t_0 と表記される駐車開始モーメントを開始時刻として、エンドツーエンドのニューラルプランナに基づいて経路 $T_{t0} = N_\theta{}'$ (I_{t0}, S) を予測し、初期モーメント t_0 から現在モーメント t までの相対姿勢を定位システムで求めることができ、$ego_{t0\to t}$ と表記される。目標操舵角A^{tar}は、RWF(リアホイールフィードバック)法を用いて求めることができ、以下のように表すことができる：

$$\mathcal{A}^{tar}_t = \mathrm{RWF}(\mathcal{T}_{t_0}, ego_{t_0\to t}). \tag{8}$$

シャーシからの速度フィードバックV^{feed}とステアフィードバックA^{feed}、および設定からの目的速度V^{tar}と計算からの目的ステアA^{tar}に応じて、カスケードPIDコントローラを利用して横方向と縦方向の制御を実現する。新しい予測軌道が生成された後、T_{t0}と$ego_{t0\to t}$はリセットされ、車両制御プロセス全体を通してグローバルなローカライゼーションに頼る必要がなくなる。

IV. Experiments

A. データセットの収集

データセットは、車両搭載デバイスを使用して収集されている。包括的な視覚認識と軌跡を容易にするために、RGB画像を撮影するためにサラウンドビューカメラが採用されている。同時に、デッドレコニング技術も統合され、センサーデータフュージョンアルゴリズムを活用することで、ロバストで正確な車両定位を実現している。実験プラットフォームと使用したセンサーのレイアウトを図4に示す。図5に示すように、地下ガレージや地上ガレージなど、様々な駐車場のシナリオにわたってデータを収集している。多様な環境からデータを収集することで、ニューラルネットワークの汎化能力を高めることができる。

図4: 実験プラットフォームとしてチャンガン車を使用。車両は、モデルの推論と制御を実行するために、Intel NUC デバイスを利用する。

図5:システムのトレーニングとテストに、いくつかの異なるガレージが利用されている。ガレージIとIIの駐車場データの一部を学習に使用。一方、Garage I と II の残りの駐車スロットデータは、Garage III と IV の収集したすべてのスロットデータと同様に、トレーニングに関与していないものを使用してテストを行う。

B. Implement Details

学習過程において、サラウンドビューカメラ画像(カメラ台数Rは4台)を入力とし、駐車終了時のある地点により目標駐車スペースを決定する。軌跡シーケンスポイントは、エンドツーエンドの予測結果を監督するために使用される。

推論プロセスでは、RVizインターフェースソフトウェアの ”2D-Nav-Goal ”を用いて目標駐車場スロットを選択し、目標駐車場スロットを取得する。このモデルは、サラウンドビューカメラからの現在の画像とターゲットスロットを取り込み、自己回帰的に後続のn個の軌跡点の位置を予測する。コントローラは、経路計画結果、自我のポーズ、フィードバック信号に基づいて車両を操縦し、車両を指定されたスロットに駐車させる。注目すべきは、目標点と予測軌跡点の座標が車両座標フレームで表現され、軌跡シーケンスとBEV特徴が一貫した座標ベースで表現されることを保証することである。この設計はまた、システム全体をグローバル座標フレームから独立させる。

Regarding the neural network details, the size of BEV

表1：異なる駐車場シナリオにおける駐車場性能テストの定量的結果

Garage	Scene	PSR (%) ↑	NSR (%) ↓	PVR (%) ↓	APE (m) ↓	AOE (deg) ↓	APT (s) ↓	APS ↑
Garage I	Scene A	70.3	3.7	62.9	0.59	7.2	64	51.5
	Scene B	90.7	0.0	38.8	0.58	3.9	70	81.5
	Scene C	83.3	8.3	58.3	0.62	5.8	60	63.5
Garage II	Scene A	83.3	0.0	50.0	0.35	10.0	66	58.0
	Scene B	91.6	0.0	58.3	0.47	6.5	63	69.7
	Scene C	81.2	0.0	50.0	0.60	6.8	64	64.6
Garage III	Scene A	95.8	0.0	33.3	0.20	2.5	51	88.6
	Scene B	100.0	0.0	25.0	0.34	5.2	50	83.1
	Scene C	91.6	8.3	41.6	0.93	7.4	55	68.6
Garage IV	Scene A	94.3	0.0	16.7	0.50	3.0	81	82.1
	Scene B	88.7	0.0	11.1	1.14	7.4	86	75.1
	Scene C	83.3	16.6	33.2	0.56	3.5	96	65.9

特徴量は200×200であり、x∈[-10m, 10m], y∈[-10m, 10m]の実際の空間範囲に0.1mの分解能で対応している。変換デコーダでは、軌跡直列化N_tの最大値は1200である。軌跡デコーダは、長さ30の予測シーケンスを生成し、推論において精度と速度の最良のバランスを達成する。

PyTorchフレームワークを用いて本手法を実装する。ニューラルネットワークは、バッチサイズ16のNVIDIA GeForce RTX 4090 GPU1台で学習され、総学習時間は約8時間、フレーム数は40,000フレームである。テストデータは約5,000フレームからなる。

C. Evaluation Metrics

1）モデルの軌跡評価：実際のシナリオ実験を行う前にモデルの性能を分析するために、モデルの推論能力を評価するための評価指標をいくつか設計する。

L2 距離（L2 Dis.）L2 距離とは、予測された軌道と真実の軌道のウェイポイント間の平均ユークリッド距離のことである。この指標は、モデル推論の精度と正確さを評価するものである。

Hausdorff Distance（Haus. Dis.）Hausdorff Distanceとは、2つの点集合間の最小距離の最大値を指す。この指標は、点集合の観点から、予測された軌跡が真実の軌跡とどの程度一致するかを評価する。

フーリエ記述子の差(4.Diff.）フーリエ記述子 Difference は、軌跡間の差を測定するために使用することができる。値が小さいほど、軌跡間の差が小さいことを示す。この指標は、実際の軌跡と予測された軌跡の両方をベクトルとして表現するために、一定数のフーリエ記述子を使用する。

2）エンドツーエンドの実車評価：実車実験では、エンドツーエンドの駐車性能を評価するために、以下のメトリクスを使用する。

駐車場成功率(PSR）駐車場成功率は、エゴ車両が目標駐車場スロットに駐車に成功する確率を表す。

スロット率なし(NSR）指定された駐車場での駐車の失敗率。

駐車場違反率(PVR）駐車場違反率とは、車両が指定された駐車場をわずかに超え、隣接する駐車場を妨害したり、妨げたりすることがない状況を指す。

平均位置誤差(APE）平均位置誤差とは、目標駐車位置と駐車成功時の自車両の停止位置との間の平均距離のことである。

平均方位誤差(AOE）平均方位誤差とは、目標とする駐車方位と、駐車が成功したときの自車両の停止方位との差の平均値である。

平均駐車スコア(APS）平均駐車スコアは、駐車中の位置誤差、方位誤差、成功率を含む総合的な評価によって算出される。スコアは0から100の間に分布している。

平均駐車時間(APT）複数の駐車操作における平均駐車時間。駐車時間は、駐車モードが開始された瞬間から、車両が指定されたスペースに正常に駐車するまで、または異常または故障により駐車プロセスが終了するまで測定される。

D. 定量的結果

提案したエンドツーエンドの駐車システムを用いて、4つの異なる駐車場でクローズドループ車両テストを実施し、提案システムの性能を検証した。結果を表 I に示す。

実験では、4つの異なるガレージでテストを行った。ガレージIは地下ガレージ、ガレージII、III、IVは地上ガレージである。各ガレージについて、3つの異なる実験シナリオを実施した。シーンAは障害物のない駐車場である。シーンBは、左側または右側に車両がある駐車場である。シーンCは、近くに障害物や壁がある駐車場である。各実験シナリオについて、3つの異なる駐車場をランダムに選択した。各スロットの左右について、約3回の駐車テストを実施した。

図6:異なるシナリオにおける駐車プロセスの説明図。各行は駐車場のケースを示す。自動車や壁など、隣接する駐車スペースに障害物がある場合でも、本手法は車両を効果的に操縦し、指定された場所に駐車させることができる。

実験結果より、提案手法は様々なシナリオにおいて高い駐車成功率を達成し、ロバストな駐車能力を示すことが示された。

最近、よりエンドツーエンドの自律走行アプローチが出現しているにもかかわらず、そのほとんどは、都市走行シナリオで遭遇する課題への対処に集中している。駐車場シナリオではParkPredict[20]のような手法が採用されているが、そのタスクは我々の手法とは大きく異なる。我々の知る限り、我々のアプローチと直接比較できる既存の効果的なエンドツーエンドの方法は存在しない。表IIで我々の手法(変換器ベースのデコーダ)とトランスフォーマー(GRUベースのデコーダ)の結果を比較する。変換器ベースのデコーダは、変換器の注意メカニズムにより、より良い予測精度を持つ。

表 II: 比較パフォーマンス評価

Method	Haus. Dis. (m) ↓	L2 Dis. (m) ↓	Four. Diff. ↓
Ours	**0.076**	**0.033**	**0.43**
Transfuser [3]	0.676	0.458	11.51

E. Ablation Study

異なるネットワーク設計の影響を分析するために、アブレーション実験をデザインした。ネットワーク構造に関しては、表IIIに示すように、特徴量融合に関するアブレーション実験を行った。ベースライン(ターゲットクエリ)、特徴連結、特徴要素別加算の結果を比較する。ターゲットクエリアプローチは、ターゲット特徴とBEV特徴を完全に統合するために、注意と空間アライメントメカニズムを利用する。ターゲットスロットとBEV画像との空間的な関係を明示的に制約し、最も高い軌跡予測精度を達成する。

F. Visualization

異なるシナリオにおける駐車プロセスを図6に示すが、本アルゴリズムの多様なシナリオにおいて、汎用性の高い適応能力を示している。

表III:特徴量融合に関するアブレーション研究

Method	Haus. Dis. (m) ↓	L2 Dis. (m) ↓	Four. Diff. ↓
Baseline	**0.076**	**0.033**	**0.43**
Concatenation	0.098	0.045	0.79
Element-wise	0.097	0.047	0.83

G. Limitations

提案手法は駐車タスクにおいて優位性を示すが、まだいくつかの限界がある。第一に、本手法はデータ規模やシナリオの多様性に制約があるため、移動目標に対する適応性が低い。データセットを拡張することで、移動物体に対するモデルの適応性を向上させることができる。第二に、エキスパートの軌跡を利用した学習プロセスのため、効果的なネガティブサンプルを提供することは不可能である。さらに、駐車中に大きな乖離が生じ、最終的に駐車の失敗につながる場合、強固な是正メカニズムは存在しない。その後、NeRF[31](Neural Radiance Field)と3DGS[32](3D Gaussian Splatting)を利用することで、実世界の条件に近いシミュレータを構築することで、深層強化学習を用いてエンドツーエンドモデルを学習することができる。最後に、我々のエンドツーエンドの駐車方法は良好な結果を得ているが、従来のルールベースの駐車方法と比較すると、依然としてギャップがある。しかし、エンドツーエンドの技術が進歩し続けるにつれて、この問題は解決されると考えている。エンドツーエンドの駐車アルゴリズムが、今後複雑なシナリオで優位性を示すことを期待している。

V. Conclusion

本論文では、カメラベースのエンドツーエンドの駐車モデルを提案する。このモデルは、ターゲットスロットとサラウンドビューのRGB画像を入力し、ターゲットクエリによってBEVビューの融合特徴を取得し、自己回帰的に変換デコーダを用いて軌跡点を予測する。軌跡計画の結果は、その後、制御のために利用される。

提案手法を様々なシナリオで広範囲に評価し、その信頼性と一般化可能性を示した。とはいえ、我々のエンド・ツー・エンドの手法と高度に最適化されたルールベースの駐車手法との間には、まだ性能差が存在する。今後の課題として、学習ベースのアプローチが最終的に従来の手法を凌駕することを期待し、エンドツーエンドの駐車アルゴリズムの性能をさらに向上させることを目指す。私たちは、私たちの研究と実践が、研究者やエンジニアの仲間にインスピレーションを与え、思い起こさせると信じている。

REFERENCES

[1] T. Qin, T. Chen, Y. Chen, and Q. Su, "AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5939–5945, IEEE, 2020.

[2] Y. Yang, D. Chen, T. Qin, X. Mu, C. Xu, and M. Yang, "E2e parking: Autonomous parking by the end-to-end neural network on the carla simulator," in *Conference on IEEE Intelligent Vehicles Symposium*, 2024.

[3] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*, pp. 66–75, PMLR, 2020.

[5] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

[6] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6119–6132, 2022.

[7] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

[8] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Conference on robot learning*, pp. 1–16, PMLR, 2017.

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

[10] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," *IEEE Transactions on Intelligent Vehicles*, 2023.

[11] M. Bansal, A. Krizhevsky, and A. Ogale, "ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst," 2018.

[12] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7073–7083, 2021.

[13] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Proceedings of The 6th Conference on Robot Learning*, vol. 205 of *Proceedings of Machine Learning Research*, pp. 726–737, PMLR, 14–18 Dec 2023.

[14] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural attention fields for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15793–15803, 2021.

[15] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 4693–4700, IEEE, 2018.

[16] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

[17] 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*, pp. 14403–14412, 2021.

[18] S. Rathour, V. John, M. Nithilan, and S. Mita, "Vision and dead reckoning-based end-to-end parking for autonomous vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2182–2187, IEEE, 2018.

[19] R. Li, W. Wang, Y. Chen, S. Srinivasan, and V. N. Krovi, "An end-to-end fully automatic bay parking approach for autonomous vehicles," in *Dynamic Systems and Control Conference*, vol. 51906, p. V002T15A004, American Society of Mechanical Engineers, 2018.

[20] X. Shen, I. Batkovic, V. Govindarajan, P. Falcone, T. Darrell, and F. Borrelli, "Parkpredict: Motion and intent prediction of vehicles in parking lots," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1170–1175, IEEE, 2020.

[21] X. Shen, M. Lacayo, N. Guggilla, and F. Borrelli, "Parkpredict+: Multimodal intent and motion prediction for vehicles in parking lots with cnn and transformer," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3999–4004, IEEE, 2022.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 09–15 Jun 2019.

[23] J. Philion and S. 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*, pp. 194–210, Springer, 2020.

[24] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. E. Hinton, "Pix2seq: A language modeling framework for object detection," *ArXiv*, vol. abs/2109.10852, 2021.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[26] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Conference on Robot Learning*, pp. 180–191, PMLR, 2022.

[27] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in *European Conference on Computer Vision*, pp. 531–548, Springer, 2022.

[28] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*, pp. 1–18, Springer, 2022.

[29] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1477–1485, 2023.

[30] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "BEVFusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10421–10434, 2022.

[31] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[32] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.