E2Eパーキング: CARLAシミュレータ上のエンドツーエンドニューラルネットワークによる自律的なパーキング

Yunfan Yang, Denglong Chen, Tong Qin*, Xiangru Mu, Chunjing Xu, and Ming Yang

概要- 自律駐車は、特に混雑した駐車場において、インテリジェントな車両にとって重要なアプリケーションである。閉鎖空間は、非常に正確な知覚、計画、制御を必要とする。現在、幾何学的な知覚とルールに基づく計画を利用する従来の自動駐車支援(APA)システムは、単純なシナリオで駐車タスクを支援することができる。ノイズの多い測定では、手作りルールは様々な環境において柔軟性と頑健性に欠けることが多く、超混雑した狭い空間では性能が低い。逆に、経験豊富な人間のドライバーは多く、明示的なモデリングやプランニングをすることなく、狭いスロットに駐車するのに適している。このことに触発され、ニューラルネットワークは、手作りのルールなしに、専門家から直接公園する方法を学習することが期待される。そこで、本論文では、駐車タスクを処理するためのエンドツーエンドのニューラルネットワークを紹介する。入力は周囲のカメラで撮影された画像と基本的な車両の動作状態であり、出力はステア角、加速度、ギアなどの制御信号である。ネットワークは、経験豊富なドライバーを模倣することで、車両を制御する方法を学習する。CARLAシミュレータで閉ループ実験を行い、駐車タスクにおいて提案するニューラルネットワークによる車両制御の実現可能性を検証した。実験により、我々のエンドツーエンドシステムが、0.3mと0.9度の平均位置誤差と方位誤差を達成し、全体の成功率は91%であることが実証された。コードはhttps://github.com/qintonguav/e2e-parking-carlaで入手可能。

(a) 自律駐車タスク

(b) Overall Structure

図1:全体構造の説明図。我々は、ニューラルネットワークによって視覚情報を制御信号に直接マッピングする、自律駐車のためのエンドツーエンドのフレームワークを提案する。ニューラルネットワークは、専門家を模倣して車両を制御することを学習する。動画デモは https://youtu.be/SfnVl3YdQow にある。

I. Introduction

自律走行の出現は、安全に航行し駐車できる高度なシステムの開発に重点を置き、運輸業界に革命をもたらした。特に自律駐車は、狭い空間での複雑な操縦と精密な制御を必要とするため、一連の課題を提起する。自律駐車に対する従来のアプローチ[1]は、知覚、マッピング、計画、制御モジュールを含む多段階のパイプラインを含むことが多く、それぞれに複雑さと限界がある。センサーのノイズとモデルの不確実性により、システムは段階的に誤差を蓄積する。さらに、計画や制御における手作りのルールは柔軟性に欠け、蓄積されたエラーに悩まされやすい。したがって、従来の多段ルールベースシステムは壊れやすく、様々な複雑な環境でのスケールアップに失敗している。

近年、自律走行における複雑なタスクに取り組むための有望な代替手段として、エンドツーエンドの学習パラダイムが登場している。

Yunfan Yang、Denglong Chen、Xiangru Mu、Chunjing Xuは、中国上海市のHuawei TechnologiesのIAS BUに所属している。秦同、楊明は、上海交通大学未来技術局(中国・上海)所属。{yangyunfan8, chendenglong, muxiangru, xuchunjing}@huawei.com, {qintong, mingyang}@sjtu.edu.cn.*は対応する著者である。

生のセンサー入力を制御アクションに直接マッピングすることで、エンドツーエンドシステムは、パイプライン全体を簡素化し、適応性を高め、手作りの特徴やルールへの依存を減らす可能性を提供する。このアプローチは、車線維持、物体検出、衝突回避などのタスクで顕著な成功を示している。エンドツーエンド学習の利点に動機づけられ、我々はこのパラダイムの新しい応用例であるE2Eパーキングを紹介する。

我々のアプローチでは、入力は車載カメラで撮影されたサラウンド画像と基本的な車両運動状態であり、出力はステア角、加速度、ギアからなる車両シャーシ制御信号である。このエンドツーエンドのフォーマットを採用することで、明示的に手作りされたモデルやルールなしに、車両がスムーズかつ効率的に駐車操作を実行できるようにすることを目指す。さらに、外部ローカライゼーションモジュールを必要としないシステムを設計することを提案する。車両は、駐車目標が最初に選択された時点で追跡することができる。エンドツーエンドのシステムは、広範な訓練と微調整を通じて、視覚的な合図から制御動作への複雑なマッピングを学習し、車両が自律的にナビゲートし、駐車枠に駐車することを可能にする。提案手法の有効性を評価するため、定量的な指標を設計し、CARLA [2]シミュレータで閉ループ実験を行った。

その結果、本手法は、平均位置・姿勢誤差0.3m、0.9度で、総合成功率91%に達することができた。今後の研究者の便宜のため、CARLA駐車場データセットとソースコードを公開する予定である。全体として、本研究の主な貢献は以下の通りである：

- 画像と運動状態を駐車の制御信号に直接マッピングする、新規かつ実現可能なエンドツーエンドの視覚的駐車ソリューションを提案した。手作りの特徴やルールの必要性を排除することで、本システムはパイプライン全体を簡素化し、多様な駐車シナリオへの適応性を向上させる。
- CARLAシミュレータ上でクローズドループ実験を行い、エンドツーエンドの自律駐車システムの実現可能性と信頼性を検証するための包括的な評価指標を提案した。
- CARLAにおける駐車タスクの定量的ベンチマークを確立し、CARLAで生成された駐車データセットを公開した。

II. II. 文献レビュー

過去数十年にわたり、ニューラルネットワークは大きなブレークスルーを遂げてきた。まず、ニューラルネットワークでよく使われるコンポーネントやアーキテクチャ、例えば変換器やBEVモジュールを見落としてしまう。次に、これらのニューラルネットワークをどのように自律走行タスクに採用するかをレビューする。

A. 変換モデル

Transformer[3]はもともと自然言語処理(NLP)の分野で応用され、長距離依存性を捉えるための独自の自己注意メカニズムを活用することで、GPT[4]-[6]のようなシーケンスモデリングタスクで大きな成功を収めている。連続的な探索を通して、変換器アーキテクチャはコンピュータビジョンや自律走行タスクの広範な応用にも使用されている。コンピュータビジョンのために、ViT [7]は画像をパッチに分割し、それらを逐次入力として符号化することで、変換器が大域的・局所的な視覚的特徴の両方を捉えることを可能にし、画像分類性能を向上させた。DETR[8]は、アテンション機構を採用することで、エンドツーエンドのプロセスを通じて正確な物体検出を実現する、変換器ベースの物体検出アプローチである。自律走行のために、Transfuser [9]とInterFuser [10]は、画像とLiDARからの特徴のマルチスケール融合のための変換器を設計した。UniAD[11]は、変換デコーダに基づく知覚・予測モジュールを含む、統一された自律走行フレームワークを構築した。CIL++[12]は、マルチビュー画像、速度、ナビゲーションコマンドから特徴を集約するために変換エンコーダを使用する、エンドツーエンドの自律走行フレームワークを提案した。ParkPredict+ [13]は、駐車場のウェイポイントを予測するために変換器を使用した。上記の研究に触発され、我々は駐車シナリオにおける制御信号を直接出力するために、変換器のフレームワークを採用する。

B. BEV表現

鳥瞰図(BEV)は、セマンティックセグメンテーションM2Bev[14]やCVT[15]、物体検出BEVDet[16]やBEVFormer[17]など、自律走行分野でよく使われるデータ表現である。障害物、車線、その他の交通参加者に関する重要な3D情報を取得し、車両の周囲環境のトップダウン視点を提供する。奥行き情報がない場合、2D画像特徴を直接BEVに変換することはできない。BEVへの2次元情報の転送方法については、数多くの研究がなされている。LSS [18]は、各画素の深度分布を学習し、カメラパラメータを用いてフラストレーションをBEV表現に変換した。BEVDepth[19]は、LSSに基づく明示的な深度監視を追加し、深度推定精度を向上させた。Transfuserとその亜種[9, 20]は、融合されたLidarと画像特徴からBEV表現を生成するために変換器を採用した。ThinkTwice[21]は、LSSに依存して、その後のデコードのために中間BEV特徴を抽出した。ST-P3[22]は、過去数回のタイムスタンプにおいて、マルチビューカメラから時空間BEV特徴を学習した。精度と効率のバランスをとるため、LSS[18]で使用されているモジュールを採用し、BEV表現を生成する。

C. エンドツーエンドの自律走行

近年、ニューラルネットワークに基づくエンドツーエンドの自律走行[23]が話題になっている。従来の階層的アプローチ[24]と比較して、エンドツーエンドの手法は重いルールに依存せず、様々なシナリオに対してより強力な汎化能力を持つ。

都市シナリオ Wayve[25]とOpenpilot[26]は、カメラ画像から直接ウェイポイントを予測するエンドツーエンドのニューラルネットワークを提案した。BEV特徴からの自動回帰ウェイポイント生成には、Transfuser [9]、Transfuser++ [20]、InterFuser [10]など、多くの手法がGRU [27]を採用している。上記のウェイポイントを直接出力する方法では、最終的なアクションを達成するためにコントローラ(PIDなど)が必要であった。しかし、このコントローラは、いくつかの複雑なシナリオでトラッキングエラーが発生している。CIL[28]は、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直接マッピングして信号を制御するニューラルネットワークを提案した。学習データの偏りや因果関係の混乱を軽減するために、CILRS [29]はResNetバックボーンとCILに基づく速度予測ブランチを追加した。LSD [30]は、複数の運転シナリオに適応するために、複数の戦略を持つハイブリッド模倣学習モデルを設計した。TCP [31]はウェイポイントGRUブランチと制御GRUブランチを融合し、性能を向上させた。ThinkTwice[21]は、粗いものから細かいものへの戦略によってウェイポイントを生成し、信号を制御するためのスケーラブルなデコーダモジュールを設計した。

駐車場のシナリオ: 現在、駐車場タスクのためのいくつかの方法は、伝統的な階層的フレームワーク [32, 33]に焦点を当てているが、エンドツーエンドのニューラルネットワーク駐車場の方法はほとんど登場していない。Rathourら[34]は、前後の画像をステアリングとギアに直接マッピングするCNNモデルを設計した。Liら[35]は、CNNへの入力としてリア画像のみを使用し、ステアの角度と速度を出力した。ParkPredict[36]は、CNNとLSTMに基づく駐車場とウェイポイント予測ネットワークを提案したが、入力には生のカメラ画像も含まれていなかった。ParkPredict+[13]はParkPredictのLSTMを変換器に置き換えたもので、入力は意味的なBEV画像で構成されている。

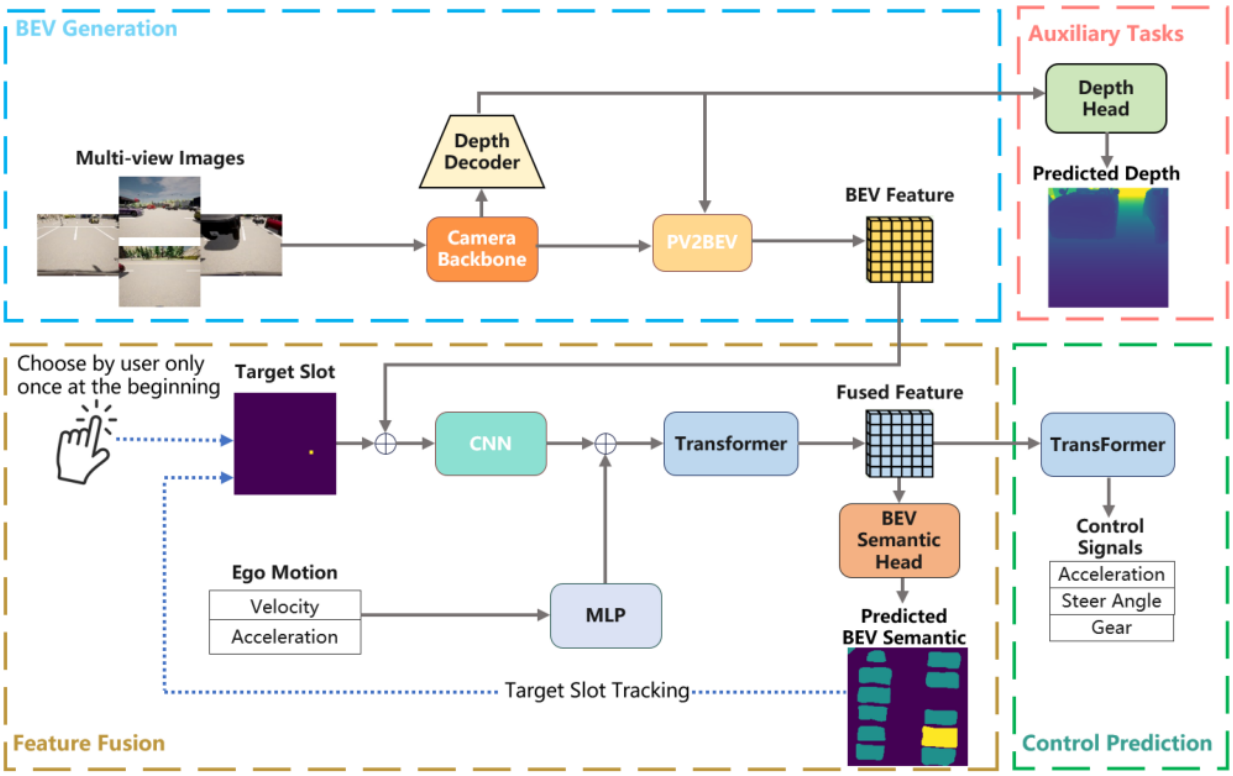


図2:提案システムのフレームワークは、BEV生成、特徴量融合、制御予測、補助タスクを含む。BEV生成モジュールは、LSS[18]法を用いて、周辺画像をBEV特徴マップに変換する。次に、BEV特徴量、車両運動量、ターゲットスロットを変換器を介して融合する。BEVグリッド(黄色の点)内のターゲットスロットの相対位置は、余分なチャネル上に描画され、BEV特徴マップに連結される。言語モデリングに基づく変換デコーダは、融合された特徴を受け取り、制御信号を予測する。深度予測は補助タスクとして教師される。BEVの意味予測における緑色の四角は周囲の車両であり、黄色の四角は目標駐車場である。

上記の方法はいずれも、4枚の生画像から完全な制御信号(ステア角、加速度、ギア)を直接得ることはできない。エンドツーエンドの形式を最大限に活用するために、視覚画像を直接制御信号にマッピングする変換器を使用する。

III. METHODOLOGY

本研究では、視覚的なエンドツーエンドの駐車のための新しいアーキテクチャを提案する。自律駐車タスクに対処するための従来のアプローチは、通常、複数の独立したモジュールから構成される多段パイプラインに依存している。しかし、我々のアプローチは、これらのモジュールを、視覚的測定値を車両制御信号に直接マッピングする1つの統一されたネットワークに統合する。

A. 問題の定式化

センサーの測定値が与えられた場合、目標は車両を安全かつ正確に所望の駐車スロットに制御することである。モデルは、D : τ^n, n∈ (0, N) のデータセットで教師あり学習され、エキスパートからのN個の駐車軌跡を含む。各軌跡は、各瞬間の観測信号と制御信号を含む: τ^n : [X_t^n, C_t^n], t∈ (0, T^n).X_t^nは、時刻tにおけるサラウンドカメラからの生画像、ターゲットスポット位置、n番目のルートにおける車両の速度を含む。同様に、C_t^nは時刻tにおけるn番目のルートでの加速度、操舵角、ギア状態を含む。入力をX、出力をCとする教師ありの学習を考える。観測値を制御信号に対応付けるための政策πを学習する。したがって、システムは次のように定式化できる:

$$\arg\min_{\pi} \mathbb{E}_{(\mathcal{X},\mathcal{C})\sim\mathcal{D}}[\mathcal{L}(\mathcal{C}, \pi(\mathcal{X}))], \tag{1}$$

ここで、L は損失関数である。ニューラルネットワークの解釈可能性を高めるために、さらに深度予測とBEVセグメンテーションタスクを追加する。III-Dで詳しく説明する。

B. Input and Output

入力表現: カメラ画像、車両の状態、ターゲット駐車場の位置をモデルへの入力として使用する。駐車場の線が撮影できる位置に4台のカメラ(前面、左面、右面、後面)を設置する。車両の状態には、車輪走行距離計とIMU(慣性計測ユニット)から得られる自車両の速度と加速度が含まれている。駐車場のシナリオでスムーズで合理的な制御信号を生成するためには、モデルが現状を認識する必要がある。駐車場として、ターゲットスロットが必要である。

ターゲットスロットは、車両のBEV座標の下のポイントを指し示すユーザーによって初めて指定される。以下では、セグメンテーションタスクにより、ターゲットスロットを連続的に追跡する。

出力表現：我々のモデルは、0.1秒の時間間隔で次の4ステップの車両制御信号を予測する。正規化加速度$acc_t \in [-1, 1]$、操舵角、操舵$_t \in [-1, 1]$、ギア(前方、後方)、ギア$_t \in \{0, 1\}$を予測する。制御モジュールでは、加速度はスロットルとブレーキに線形にマッピングされ、ステアリング値はステアリングホイール角度に線形にマッピングされる。Pix2seq[37]に触発され、多段階の制御信号予測を言語モデリングタスクとする。このモデルは、文を形成するためにすべての単語を順次生成するのと同様に、制御信号を一つずつ出力する。制御信号をトークン化するために、加速度とステア角は0.01の分解能で離散化される。したがって、加速度とステアは[0, 200]の離散値となる。加速度については、0は完全なブレーキ、200は完全なスロットルを示す。ステアリングの0は左端、200は右端を意味する。開始(BOS、シーケンス開始)と終了(EOS、シーケンス終了)トークンを追加することで、シーケンスは以下のように記述できる：

$$\begin{aligned} S &= [BOS, \mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, EOS] \\ \mathbf{c}_n &= [acc_n, steer_n, gear_n]. \end{aligned} \tag{2}$$

多段階予測により、このモデルが短期間で滑らかで一貫性のある出力を生成できることを期待している。

C. ネットワークアーキテクチャ

図2に示すように、我々の提案するアーキテクチャは、いくつかのサブモジュールから構成される：BEV生成、特徴融合、制御予測である。以下では、それぞれのセクションの詳細を説明する。

1) BEVの生成：カメラビューからBEVへの変換はLSS法[18]に従う。マルチビュー画像$I \in_R^{3\times H\times W}$は、まずバックボーンによって処理され、画像特徴$F_{img} \in_R^{Ci\times Hi\times Wi}$を取得する。ここで、(H, W)，($H_i$, W_i) はそれぞれ画像と画像特徴の高さと幅、C_i は特徴チャンネルである。各画像の特徴フラストレーションは、深度分布と画像特徴のドット積$_R^{D\times Ci\times Hi\times Wi}$によって作成される。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、特徴量のフラストレーションがBEVボクセルグリッドに投影される。ボクセル和は、同じボクセルの特徴をマージするために実装される。最終的なBEV特徴マップは、$F_{bev} \in_R^{Cb\times Xb\times Yb}$ と表され、(X_b, Y_b)はBEVグリッドサイズ、C_bはチャンネル数である。

2) 特徴の融合：BEV特徴量から得られる周辺視覚情報の他に、自我運動と目標スロットも必要である。特徴融合サブモジュールでは、BEV特徴マップ、ターゲットスロット、エゴモーションを1つの統一された特徴に融合する。まず、ターゲットスロットを余分なBEVグリッド$F_{target} \in_R^{1\times Xb\times Yb}$に投影し、ターゲットスロットの位置を1とし、他のグリッドを0とする。F_{bev}とF_{target}を連結することで、最終的なbev特徴量$F_{bev} \in$_{R}^{(Cb+1)×Xb ×Yb}が得られる。

2次元C′ ×L畳み込みと再形成により、C'_bチャンネル、長さLの特徴量$F_{bev}{}'_R$ bが得られる。次に、現在の自我運動(速度、加速度)をMLPモーションエンコーダで$F_{ego} \in_R^{2\times L}$と表現する。最後に、$F_{bev}{}'$ とF_{ego} (C′ +2)×Lを連結して、融合特徴量$F_{fuse} \in_R$ bを得る。特徴量F_{fuse}をTransformerエンコーダに取り込むことで、自己注意メカニズムを利用する。BEV特徴量の注目度を可視化したものをFig.5に示す。

3) 制御予測：言語モデリングに基づく変換デコーダを用いて、制御信号列を自動回帰的に予測する。注意による大域的な受容野を活用することで、変換デコーダは融合特徴量F_{fuse}からの空間情報と、シーケンス位置埋め込みからの時間情報を取り込んで制御信号を生成する。デコーダは、知覚された環境特徴から車両制御への直接的なマッピングを行う。

詳細には、融合された特徴量と埋め込まれたトークンは、クロスアテンションメカニズムを介して互いに相関する。注意のキーKと値Vは特徴融合サブモジュールF_{fuse}の出力から得られ、空のシーケンスはクエリQとして機能した。デコーダは制御信号トークンを繰り返し生成する。新たに生成されたトークンはシーケンスにスプライスされ、デコーダは新しいシーケンスを受け取り、最大長に達するまで次のトークンを生成する。

出力トークンは、正規化加速度、正規化ステア角、ギアにデトークナイズされる。次に、スロットル、ブレーク、ステア角、ギアの状態に線形マッピングする。

D. 損失と補助タスク

主な損失関数が2つ、補助タスクが1つある。1) 制御信号損失：クロスエントロピー損失を用いて、予測された制御シーケンスとグランドトゥルースの制御シーケンスとの間のギャップを測定する。

2) セマンティックセグメンテーションロス：セマンティクスの場合、モデルはBEVグリッドオブジェクトを車両、ターゲットスロット、背景の3つのカテゴリに分類する。BEVセマンティックヘッドは、融合特徴量F_{fuse}の上で動作する。意味出力もクロスエントロピー損失で教師される。

予測された意味的ターゲットスロットを用いて、ターゲットスロットのトラッキングを実現する。詳細には、予測されたターゲットスロットの中心(x,yの平均)を抽出し、次回のターゲットスロット入力として扱う。このように、ユーザは初めてターゲットスロットを任命すればよい。ネットワークはこのターゲットスロットを自動的に追跡する。訓練時には、追跡プロセスを模倣するために、ノイズの多いターゲットスロットの位置をネットワークに与える。

3) 補助タスク BEVDepth[19]は、明示的な深度監視がLSS法に大幅な性能向上をもたらすことを実証した。エンドツーエンドの手法として、深度予測を追加することで、解釈可能性も向上させることができる。そこで、補助タスクとして深度予測を追加する。グランドトゥルースの深度は、深度分布予測を監督するために、離散的な深度ビンのワンホットエンコーディングに変換される。

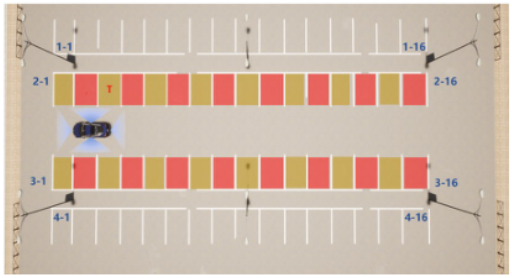


図3:駐車場の上面図。赤い駐車場はトレーニングに、茶色の駐車場はクローズドループの評価に使用される。テキストはスロットのインデックスを示す。駐車場の赤い「T」テキストは、ランダムに配置された目標駐車場を示す。

IV. Experiments

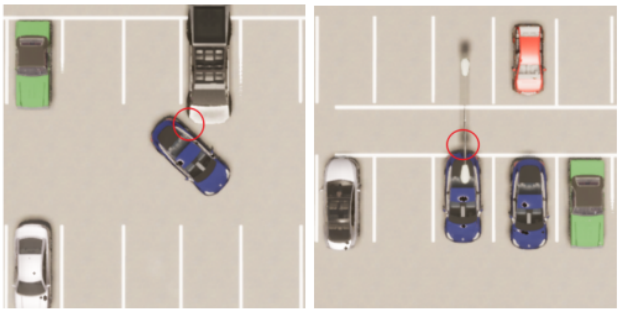
A. 実装の詳細

データ収集とクローズドループ実験にはCARLA 0.9.11シミュレータを使用した。シナリオは、図3に示すように、64の駐車スペースを含むマップTown04-Optに提供された駐車場である。

ネットワークの詳細は、BEV特徴量のサイズを200×200とし、x∈[-10m, 10m], y∈[-10m, 10m]の実際の空間範囲に対応し、分解能は0.1mである。画像バックボーンには、事前に学習させたEfficientNet-B4を用いる。深度マップの深度範囲は[0.5m, 12.5m]で、分解能は0.25mである。ネットワークの変換器構造であるエンコーダ(特徴量融合)とデコーダ(制御予測)は、共に4つの注意層と6つの注意ヘッドを持つ。
実験中、車速の過大は駐車性能に悪影響を及ぼすことが観察された。したがって、自車両の速度は、前進が12km/h、後退が10km/hに制限した。このトリックの有効性は、アブレーション研究、Sect. IV-E.
Adamのオプティマイザを利用し、重みの減衰は$1e^{-4}$、Adamのベータ値は0.9と0.999に設定した。本手法はPyTorchフレームワークを用いて実装した。ネットワークはNVIDIA Tesla V100 GPUを用い、バッチサイズ12で学習し、学習には約96時間、150エポック、22Kフレームのデータを使用した。

B. データセット収集

学習セットデータを収集するために、CARLAに基づく駐車場データ収集パイプラインを開発した。駐車場の上面図を図3に示すが、4列の駐車場と1列あたり16個の駐車場から構成されている。対象駐車場には、地面に赤い「T」記号が付けられている。各シーンにおいて、すべての駐車スロットは空のままか、ランダムな車両を割り当て、ターゲットスロットは常に占有しない。その駐車場に設置されたCARLAのプリセットされた静止車両は取り除かれる。次に、エゴ車両は、ターゲットスロットまでの距離が7m以内である限り、ターゲット駐車場スロット外の経路のランダムな位置で初期化される。自車両のヘディングは経路と平行である。



(a) 車両との衝突　　(b) ストリートランプとの衝突

図4: 衝突確率の高い2つのシーンの例。ターゲットスロットの横にある大型トラック(サイバートラック)とストリートランプは、自車両がスロットにバッキングしているときに衝突を引き起こす可能性が高い。

経験豊富なドライバーは、キーボードによって目標スロットに駐車するエゴ車両を制御する。
各収集プロセスにおいて、カメラ画像、対応する深度画像、自車両運動状態、BEVセマンティックマップを収集した。グランドトゥルースのBEVセマンティックマップは、車両の3Dバウンディングボックスを BEV平面に投影することで生成される。グランドトゥルースの深度を取得するために、データセット生成時にエゴビークルはRGBカメラと同じ位置に深度カメラを搭載する。また、制御信号はネットワーク学習用に完全に保存される。

望ましくないルートや大きな偏差がある駐車場の結果は破棄された。具体的には、自車両の最終駐車位置と所望の駐車位置との偏差は、距離が0.5m以下、向き(ヨー角)が0.5度以下でなければならない。経験豊富な4人のドライバーにより、ランダムなシーンを持つ合計128の駐車データを収集した。データサンプリング周波数は10Hzに設定されている。学習用に22Kフレームを収集した。今後の研究のヒントとして、データセットとCARLA駐車場データ生成ツールを公開する。

C. Metrics

モデルの性能を定量的に評価するために、16のパークスロットと自動運転車を各ターゲットスロットに対して24回駐車させた。評価に使用した9つのメトリクスを以下に示す。

目標成功率(TSR) 自車両が目標駐車枠にうまく駐車する確率。成功する公園と見なすには、自車両の中心から目標スロットの中心までの距離を水平方向0.6m以下、縦方向1m以下とし、方位差を10度以下とする。
目標故障率(TFR) 自車両が目標駐車場に駐車するが、その誤差が許容できない確率。
非目標レート(NTR) 自車両が非目標駐車枠に駐車する確率。

衝突率(CR) 駐車場の作業中に衝突が起こる確率。

TaskIdx	TSR (%) ↑	TFR (%) ↓	NTR (%) ↓	CR (%) ↓	TR (%) ↓	APE (m) ↓	AOE (deg) ↓	APT (s) ↓
2-1	95.83	0.00	0.00	0.00	4.17	0.23	2.23	15.49
2-3	100.00	0.00	0.00	0.00	0.00	0.29	0.37	15.38
2-5	100.00	0.00	0.00	0.00	0.00	0.35	0.48	16.13
2-7	83.33	12.50	0.00	4.17	0.00	0.35	1.07	15.90
2-9	70.83	0.00	8.33	20.83	0.00	0.34	0.44	14.50
2-11	100.00	0.00	0.00	0.00	0.00	0.31	0.67	14.68
2-13	91.67	8.33	0.00	0.00	0.00	0.32	0.53	14.22
2-15	100.00	0.00	0.00	0.00	0.00	0.18	0.55	14.53
3-1	79.17	0.00	4.17	4.17	12.50	0.28	1.11	17.53
3-3	100.00	0.00	0.00	0.00	0.00	0.22	1.25	15.61
3-5	100.00	0.00	0.00	0.00	0.00	0.42	0.60	15.68
3-7	100.00	0.00	0.00	0.00	0.00	0.24	0.37	16.39
3-9	45.83	8.33	41.67	4.17	0.00	0.36	0.83	15.00
3-11	95.83	4.17	0.00	0.00	0.00	0.40	1.38	16.75
3-13	100.00	0.00	0.00	0.00	0.00	0.33	1.12	17.98
3-15	100.00	0.00	0.00	0.00	0.00	0.25	0.91	15.70
Avg	**91.41**	**2.08**	**3.39**	**2.08**	**1.04**	**0.30**	**0.87**	**15.72**

表1: 閉ループの結果。閉ループ実験は、16の異なるシーンにおける384のテストケースから構成される。各シーンは、異なる車両開始位置で24回繰り返される。すべてのシーンはランダムに生成され、学習データとは異なる。

タイムアウト率(TR) エゴ・ビークルが指定された時間内に正常に駐車できない確率。
平均位置誤差(APE) 駐車に成功したケースにおける、エゴ車両の最終位置と目標スロットの中心との間の平均誤差。

平均方位誤差(AOE) 成功した駐車ケースにおける、自車両の最終方位(ヨー角)と所望の目標スロット方位との平均誤差。
平均駐車時間(APT) 平均駐車成功時間。

平均推論時間(AIT) ネットワーク推論の1ステップの平均時間。

D. Results

閉ループの結果 CARLAシミュレータを用いた閉ループ実験により、モデルの汎化能力を評価した。各評価シーンは、ランダムな近傍とランダムなターゲットスロットで生成された。クローズドループ評価のシーンがトレーニングデータセットのシーンと異なることを保証するために、異なるランダムシードが利用された。自車両の初期位置は、トレーニングデータセットと同じ設定であった。車両は指定された時間(30秒)以内に駐車タスクを完了する必要があった。

表 I に閉ループ実験の結果を示す。ランダムに生成された合計384のテストケースにおいて、ネットワークは全体として91.41%のTSR(目標成功率)に達し、我々のエンドツーエンド手法の有効性が実証された。16シーンのうち、ネットワークはTSR100%で9シーンを完了した。評価シーンはネットワーク学習に用いたシーンと異なるため、ネットワークが汎化能力を獲得することが示された。
ほとんどの場合、車両は所望の駐車枠まで追跡・駐車することができ、NTRはわずか2.08%(非目標率)であった。ネットワークは、最初にユーザが1回入力したものから、ターゲットスロットの位置と形状を予測することに成功した。

タスクインデックス	TSR (%) ↑	TFR (%) ↓	CR (%) ↓	APE (m) ↓	AOE (deg) ↓	APT (s) ↓
baseline	91.41	2.08	2.08	0.30	0.87	15.72
Expert	**100.00**	**0.00**	**0.00**	**0.23**	**0.48**	**14.96**
Rookie	75.00	18.75	6.25	0.35	4.00	20.13

表II:エキスパートドライバーとルーキードライバーとの比較。

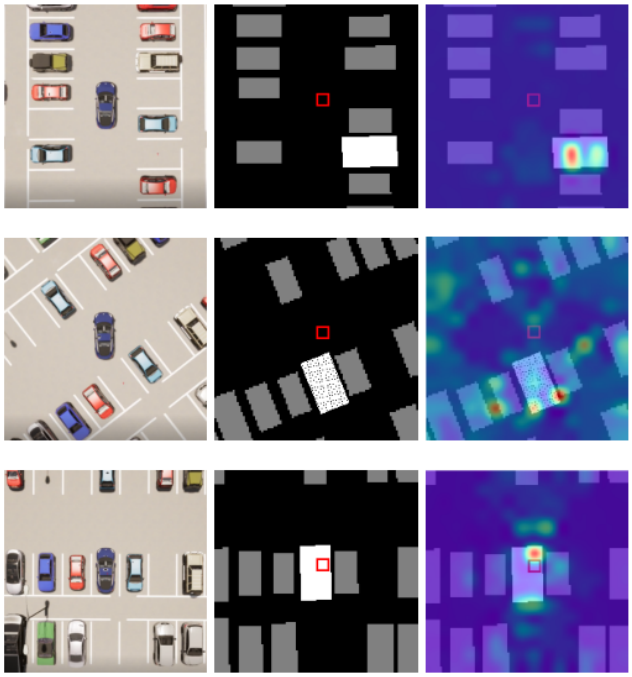
しかし、失敗事例のごく一部が2.08%のCR(Collision Rate)に寄与している。我々は、大きすぎるトラックや、ターゲットスロット付近の未見物体(街灯)が、自車両と衝突する可能性が高いことを観察した。同様のシーンを図4に示す。シーンから大型車両を取り除いた場合、CRは0.5%程度まで低下する可能性がある。トラックや街灯のシーンを増やして学習データセットを増やすことで、この問題を軽減できる可能性がある。
平均駐車位置誤差は0.3m、方位誤差は0.87°であった。図5に示すように、変換機構におけるBEVの注意を可視化した。注意はターゲットスロットに最も集中した。具体的には、エゴ車両が停止しようとしているときに、目標駐車場の停止線(リア)に注目した。このことから、ネットワークは暗黙のうちに停止線検出に依存して、いつ車両を完全に停止させるかを決めていることがわかった。

エキスパートやルーキーと比較したパーキング性能。本モデルによる制御性能を、エキスパートと新人ドライバーで比較する対比実験を行った。エキスパートと新人のデータは、我々の学習データと同じである。表IIに示すように、エキスパートが最も高いTSRと最も低いAPE、AOE、APTを示した。我々のモデルの結果は、これらのメトリクスの専門家よりもわずかに低かった。しかし、新人ドライバーと比較すると、我々のモデルはTSRで16ポイント高く、AOEで3度小さいことを達成した。また、新人ドライバーは、当社モデルよりも5秒多くスロットに駐車した。以上の比較から、我々の駐車モデルはエキスパートと同程度の性能を持ち、新人ドライバーを助けることができることがわかる。

ランタイム。を用いた駐車実験を行った。

TaskIdx	TSR (%) ↑	TFR (%) ↓	NTR (%) ↓	CR (%) ↓	TR (%) ↓	APE (m) ↓	AOE (deg) ↓	APT (s) ↓	AIT (ms) ↓
baseline	**91.41**	2.08	3.39	2.08	**1.04**	0.30	0.87	15.72	74.92
w/o depth	77.08	5.21	5.47	6.25	5.99	0.29	0.80	16.37	73.66
w/o speed limit	81.51	4.43	5.21	4.43	4.43	0.39	1.25	**13.17**	79.34
MLP decoder	83.33	**1.30**	**2.86**	**1.04**	11.46	**0.25**	**0.54**	16.59	**65.72**

表III:アブレーション研究。w/o深度はベースラインから深度監視を除去する。w/o速度制限は、クローズループ評価における速度制限のトリックを無効にする。MLPはデコーダを変換器からMLPに変更する。



(a) BEV Real (b) BEV Semantic (c) BEV Attention

図5:BEVにおける自己アテンションの可視化。(a)はCARLAシミュレータにおける実際のBEVシナリオである。(b)はBEVセマンティックのグランドトゥルースを示しており、灰色の矩形は静的車両、白い矩形は目標駐車スロットである。赤い四角はクエリの位置を示す。赤い四角へのアテンションマップを(c)に示す。駐車場処理中にターゲットスロットに注目することで、ネットワークの解釈可能性を高める。

Quadro RTX 5000グラフィックカード、合計384の駐車タスクを完了。ネットワークAIT(平均推論時間)は74.92ミリ秒で1ステップであるため、ネットワークは約13Hzで動作する。ネットワークの設計目標は、10Hzで制御信号を生成することである。したがって、本ネットワークの実行時性能は、自律駐車システムにとって、安全で効率的な運転を実用的に確保するために極めて重要な要件を満たしている。

E. アブレーション研究

補助タスク。表IIIでは、深さ補助タスクの効果を比較した。深度監視はBEVの性能を向上させることができる。したがって、深度監視を除去した後のTSRは約14ポイント減少した。実験結果は、深度監視の有効性を証明するものである。

スピードリミット。表IIIでは、スピードリミットのトリックの必要性について議論した。速度制限を行わない場合、TSRは約10ポイント低下した。高速走行によりAPTは低下したが、故障(TFR、NTR、CR、TR)は増加した。速度制限のトリックでは、システムはより安定していた。

デコーダのアーキテクチャ。また、異なる種類のデコーダの性能も比較した。表IIIに示すように、MLPデコーダはTSRで約83ポイントを獲得した。ベースラインであるトランスフォーマーデコーダーは、TSRで約8ポイント上昇した。この性能差は、主に変換器のクロスアテンションメカニズムの恩恵を受けており、我々のアプローチはBEV特徴を効果的に利用することができる。

制御ステップ。学習におけるデコーダの制御ステップは、閉ループの性能に影響を与える可能性がある。ベースラインは4つの制御ステップを使用した。制御ステップ数を変えてテストした: 1と6である。表IVに示すように、4ステップと6ステップはTSRで同程度の性能を示した。しかし、ステップを1に設定すると、TSRは約8ポイント低下した。その結果、ベースラインは性能と効率のバランスをより良く達成することが示された。

V. Conclusion

本論文では、自律駐車のためのエンドツーエンドフォーマットを採用し、周囲画像と動き計測を入力として利用し、制御信号を直接生成する研究を紹介した。本手法は、推論中の駐車目標を追跡することができる。我々は、視覚情報をフルに活用するために、カメラビューからBEV投影とTransformの注意メカニズムを活用した。本手法は、CARLA上で行われた閉ループ実験において、その実現可能性と有効性が証明された。

現段階では、シミュレーションに限界がある。本手法の全体的な能力は、汎化において実際の人間のドライバーの能力とまだ明確なギャップがある。また、動的な物体を扱う能力もまだ考慮されていない。将来的には、魚眼カメラや超音波レーダーのような余分な搭載センサーを利用した水平駐車や対角駐車など、より多様なシナリオで実世界実験を行う予定である。また、動的な物体との衝突をより回避し、コーナーケースを解決するために、駐車タスクに強化学習を実装する可能性を探る予定である。本論文は、自律走行システムに取り組む研究者や技術者にとって貴重なリソースとなり、セルフパーキング技術の分野におけるさらなる進歩への道を開くものであると考える。

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)*. IEEE, pp. 5939–5945.

TaskIdx	TSR (%) ↑	TFR (%) ↓	NTR (%) ↓	CR (%) ↓	TR (%) ↓	APE (m) ↓	AOE (deg) ↓	APT (s) ↓	AIT (ms) ↓
baseline (4-steps)	**91.41**	2.08	**3.39**	2.08	**1.04**	0.30	**0.87**	15.72	74.92
1-step control	83.07	1.04	6.77	2.86	6.25	**0.27**	0.94	17.68	74.99
6-steps control	90.63	**0.00**	6.51	**1.04**	1.82	**0.27**	1.06	**15.02**	**74.41**

表IV:トレーニングで使用したデコーダ制御ステップに関するアブレーション研究。推論段階では、閉ループ制御のために10hzで1ステップしか推論しない。ベースラインは、性能と効率のバランスをとる4ステップ戦略を選択する。

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

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020," *arXiv preprint arXiv:2010.11929*, 2010.

[8] 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*. Springer, 2020, pp. 213–229.

[9] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.

[10] 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.

[11] 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*, 2023, pp. 17 853–17 862.

[12] Y. Xiao, F. Codevilla, D. P. Bustamante, and A. M. Lopez, "Scaling self-supervised end-to-end driving with multi-view attention learning," *arXiv preprint arXiv:2302.03198*, 2023.

[13] 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)*. IEEE, 2022, pp. 3999–4004.

[14] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "M2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation," *arXiv preprint arXiv:2204.05088*, 2022.

[15] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 760–13 769.

[16] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[17] 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*. Springer, 2022, pp. 1–18.

[18] 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*. Springer, 2020, pp. 194–210.

[19] 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, no. 2, 2023, pp. 1477–1485.

[20] B. Jaeger, K. Chitta, and A. Geiger, "Hidden biases of end-to-end driving models," *arXiv preprint arXiv:2306.07957*, 2023.

[21] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li, "Think twice before driving: Towards scalable decoders for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 983–21 994.

[22] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 533–549.

[23] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[24] H. Fan, F. Zhu, C. Liu, L. Zhang, L. Zhuang, D. Li, W. Zhu, J. Hu, H. Li, and Q. Kong, "Baidu apollo em motion planner," *arXiv preprint arXiv:1807.08048*, 2018.

[25] J. Hawke, V. Badrinarayanan, A. Kendall *et al.*, "Reimagining an autonomous vehicle," *arXiv preprint arXiv:2108.05805*, 2021.

[26] L. Chen, T. Tang, Z. Cai, Y. Li, P. Wu, H. Li, J. Shi, J. Yan, and Y. Qiao, "Level 2 autonomous driving on a single device: Diving into the devils of openpilot," *arXiv preprint arXiv:2206.08176*, 2022.

[27] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[28] 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)*. IEEE, 2018, pp. 4693–4700.

[29] 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*, 2019, pp. 9329–9338.

[30] E. Ohn-Bar, A. Prakash, A. Behl, K. Chitta, and A. Geiger, "Learning situational driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 296–11 305.

[31] 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.

[32] X. Shen, E. L. Zhu, Y. R. Stürz, and F. Borrelli, "Collision avoidance in tightly-constrained environments without coordination: a hierarchical control approach," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2674–2680.

[33] J. Leu, Y. Wang, M. Tomizuka, and S. Di Cairano, "Autonomous vehicle parking in dynamic environments: An integrated system with prediction and motion planning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 890–10 897.

[34] 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)*. IEEE, 2018, pp. 2182–2187.

[35] 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. American Society of Mechanical Engineers, 2018, p. V002T15A004.

[36] 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)*. IEEE, 2020, pp. 1170–1175.

[37] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. Hinton, "Pix2seq: A language modeling framework for object detection," *arXiv preprint arXiv:2109.10852*, 2021.