模倣学習に関するサーベイ： アルゴリズム、最近の進展、および課題

Maryam Zare*, Parham M. Kebria, *Member, IEEE,* Abbas Khosravi, *Senior Member, IEEE,*
and Saeid Nahavandi, *Fellow, IEEE*

要旨-近年、ロボット工学と人工知能(AI)システムの発展は目覚ましいものであった。これらのシステムが進化し続けるにつれて、自律走行、空中ロボット工学、自然言語処理など、ますます複雑で非構造的な環境でも利用されるようになっている。その結果、手作業で行動をプログラミングしたり、(強化学習(RL)で行われるような)報酬関数を通して行動を定義したりすることは、非常に困難になっている。これは、このような環境では高度な柔軟性と適応性が要求されるため、すべての可能な状況を考慮できる最適なルールセットや報酬シグナルを指定することが困難になるためである。このような環境では、模倣によって専門家の行動から学ぶことがより魅力的になることが多い。模倣学習(IL)は、専門家の行動を模倣することによって、望ましい行動を学習するプロセスであり、それは実演を通して提供される。

本稿の目的は、ILの紹介とその基礎となる仮定とアプローチの概要を提供することである。また、この分野における最近の進歩や新たな研究分野についても詳しく説明している。さらに、本稿では、研究者がILに関連する共通の課題にどのように対処してきたかを議論し、今後の研究の方向性を示す。全体として、本稿の目的は、ロボット工学とAIにおけるILの成長分野への包括的なガイドを提供することである。

索引用語-模倣学習、デモンストレーションからの学習、強化学習、調査、ロボット工学

I. Introduction

従来、機械やロボットは自律行動を学習するために手動でプログラムされてきた[1]。従来の方法では、機械が行わなければならない動作や、機械が動作する環境の特徴について、専門家が具体的でハードコードされたルールを提供することが求められる。しかし、このようなルールを開発するには、かなりの時間とコーディングの専門知識が必要です[2]。あらゆる動作の面倒な手動ハードコーディングを自動化するために、学習アプローチが必要である[3]。模倣学習は、望ましい行動を示すことによって、それを教えるための手段を提供する。

Maryam Zare、Parham M. Kebria、Abbas Khosraviは、ディーキン大学知能システム研究・イノベーション研究所(IISRI)、Waurn Ponds、3216 VIC、オーストラリアに所属している(* Corresponding author: mzare@deakin.edu.au)。

Saeid Nahavandiは、ディーキン大学知能システム研究・イノベーション研究所(IISRI)(オーストラリア、3216 VIC、Waurn Ponds)、およびハーバード大学ポールソン工学・応用科学部(米国、マサチューセッツ州オールストン)に所属している。この研究は、出版される可能性のあるIEEEに提出された。

IL技術は、タスクを教える問題をデモを提供する問題にまで減らす可能性があり、その結果、明示的なプログラミングやタスク固有の報酬関数の開発が不要になる[3]。ILの概念は、人間の専門家が機械やロボットにプログラミングできなくても、望ましい動作を示すことができるという前提に基づいている。このように、ILは人間の専門家のような自律的な動作を必要とするあらゆるシステムで活用することができる[1]。

ILの主な目的は、エージェントがデモンストレーションの提供を通じてエキスパートを模倣することで、特定のタスクや行動の実行を学習できるようにすることである[4]。デモンストレーションは、観察と行動の間のマッピングを学習することで、タスクを実行する学習エージェントを訓練するために使用される。ILを利用することで、エージェントは、制約のある環境では単純な所定の行動を繰り返すことから、非構造化環境では最適な自律行動をとることに移行することができ、専門家に過度の負担を課すことなく移行することができる[2]。その結果、ILアプローチは、製造業[5]、ヘルスケア[6]、自律走行車[7]、[8]、ゲーム産業[9]など、幅広い産業に大きな利益をもたらす可能性がある。これらのアプリケーションにおいて、ILは、コーディングスキルやシステムに関する知識を持たない可能性のある主題専門家が、機械やロボットの自律行動を効率的にプログラムすることを可能にする。模倣による学習という考え方は以前からあったが、近年のコンピューティングやセンシングの成果に加え、人工知能アプリケーションの需要が高まり、ILの重要性が高まっている[10], [11]。その結果、近年、この分野の論文数は著しく増加している。

過去20年間に、ILに関する複数の調査が発表されており、それぞれがこの分野の発展のさまざまな側面に焦点が当てられている(図1)。Schaal[3]は、ヒューマノイドロボットの作成ルートとしてILに焦点を当て、ILの最初のサーベイを行った。より最近では、Osaら[1]がILに関するアルゴリズム的な視点を提供し、Husseinら[12]がILプロセスの各段階における設計オプションの包括的なレビューを提供している。最近では、Le Meroら[7]が、エンドツーエンドの自律走行システムのためのILベースの技術の包括的な概要を提供している。

ILに関する多くのサーベイが存在するにもかかわらず、急速に発展しているこの分野の最新の進歩を捉え、最新の技術状況を概観するためには、新たなサーベイが必要である。この分野への関心が高まり、多様な応用が進む中、包括的な調査は、新規参入者の重要な参考資料となるだけでなく、さまざまなユースケースの概要を提供することができるだろう。我々は、ILが常に進化する分野であり、新しいアルゴリズム、技術、アプリケーションが開発されていることを認める。

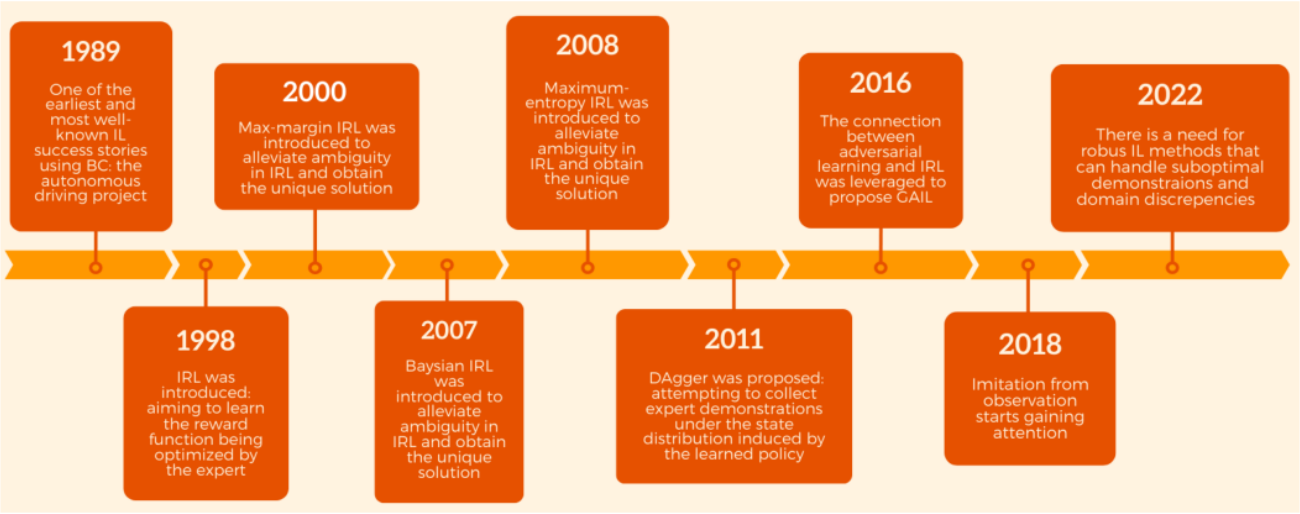

Fig. 1. A historical timeline of IL research illustrating key achievements in the field.

したがって、本調査は、ILに関する膨大な量の研究を統合し、研究者や実務家がナビゲートしやすくすることを目的としている。さらに、本研究におけるギャップと課題を明らかにし、今後の研究の方向性を明確にすることを目的とする。最後に、ILの概念と技法を、関連分野の研究者を含むより多くの読者が利用しやすくし、この分野の理解を深めることを目的とする。全体として、我々の調査はIL分野の発展に大きく貢献し、このエキサイティングな分野における今後の研究の指針になると強く信じている。

本サーベイペーパーの目的は、IL分野の包括的な概観を提示することである。そのために、歴史的・論理的な理由に基づいて、ILアプローチの議論を整理する。はじめに、ILに対する2つのアプローチ、すなわち行動クローニング(BC)と逆強化学習(IRL)を紹介する。その定式化、発展、長所、限界について述べる。さらに、敵対的模倣学習(AIL)が、学習プロセスに敵対的文脈を導入することで、どのようにIRLを拡張するかを探る。敵対的訓練をILに統合することの利点を強調し、AIL分野における現在の進歩を評価する。さらに、状態のみ(行動なし)のデモから学習することを目的とした新しい手法として、観測からの模倣(IfO)を紹介する。IfOの意義と、状態のみの観測から学習するという課題に取り組むために、これまでのBC、IRL、AILのカテゴリーをどのように取り入れ、様々な手法で拡張しているかについて議論する。最後に、IL技術が実世界のシナリオで遭遇する課題、例えば、最適でないデモンストレーションや、エキスパートと学習者の間のドメインの不一致について議論する。最後に、ILに対する様々なアプローチ、その限界、そしてそれらに対処するために取り得る将来の研究の方向性について議論する。

II. Behavioral Cloning

BCは、行動を学習する問題を教師あり学習タスクとして扱うIL技術である[13], [14]。

BCでは、環境の状態を対応する専門家の行動に対応付ける学習を行うことで、専門家の行動を模倣するモデルを学習する。専門家の行動は、デモとも呼ばれる状態アクションのペアのセットとして記録される。学習プロセスにおいて、モデルはこれらのデモンストレーションを入力として提供され、現在の状態を対応するエキスパートアクションにマッピングする関数を学習するように学習される。モデルが学習されると、学習された関数を使用して、それまで遭遇したことのない新しい状態に対するアクションを生成することができる。

BCの利点の1つは、環境の基本的なダイナミクスの知識を必要としないことである[13]。その代わりに、提供されたデモンストレーションのみに頼って動作を学習する。さらに、BCは教師あり学習モデルの学習を伴うため、計算効率が高く、機械学習における問題としてよく研究されている。

BCアプローチは単純であるにもかかわらず、共変量シフト問題[15]という大きな欠点がある。この問題は、学習時には学習者がエキスパートポリシーによって生成された状態に対して学習されるが、テスト時には学習者がその行動によって誘導された状態に対してテストされるために生じる[16]。その結果、テスト中に観測された状態分布は、トレーニング中に観測された状態分布とは異なる可能性がある。BC教師ありアプローチの問題点は、エージェントがドリフトして分布外の状態に遭遇したときに、どのようにデモされた状態に戻るかを知らないことである[17]。共変量シフトは、運転[18]のようなセーフティクリティカルな状況では特に危険である。なぜなら、エージェントは訓練中に見たことのない新しい状況に遭遇する可能性があり、ミスから回復する能力は事故を回避するために重要である可能性があるからである。共変量シフトの問題に対処し、BCアプローチの頑健性を向上させるために、3つの広範な研究領域が特定された(図2)。

最初の、そして最も人気のある分野は、対話型ILである。このタイプのアルゴリズムは、エージェントが訓練中に相談できるオンライン専門家にアクセスできるという仮定に基づいている。データセット集約(DAgger)[14]は、最も初期の対話型IL手法であり、訓練とテストの時間的ミスマッチ問題を解決するために、エージェント自身の状態分布で訓練することを提案する。

DAggerは、エージェントが収集したデータに、取るべき適切なアクションでラベル付けを変更するよう、専門家に問い合わせる。しかし、頻繁なクエリのため、人間の専門家は大きな認知的負担を受け、その結果、不正確なフィードバックや遅延フィードバックが学習プロセスに悪影響を及ぼす[19]。その結果、いつ、どのように被験者を巻き込むかを決定することは、対話型ILアルゴリズムの重要な課題の1つである[20]。継続的なフィードバックを提供するのではなく、「人間ゲート」対話型ILアルゴリズム[21]、[22]は、DAggerを拡張して、専門家が修正介入を提供するタイミングを決定できるようにする。例えば、Human-gated DAgger (HG-DAgger) [21]は、エキスパートがエージェントが状態空間の安全でない領域に到達したと判断するまで、エージェントの軌跡を展開する。この場合、人間の専門家は、システムの制御を行い、エージェントを安全な状態に戻すように導くことで介入する。この方法を用いると、人間の介入量を制限する制約がない。Liら[19]は、人間の介入を最小化し、訓練中の自動化を適応的に最大化するように学習する方法を提案している。これを達成するために、人間の専門家が介入を発行するとき、エージェントにコストがかかり、エージェントは学習プロセス中に最小化することを学習する。

しかし、これらのアルゴリズムの使用は、エージェントがいつ介入するかを決定するために、人間の専門家が常に監視しているに依存しており、そのため、人間に大きな負担がかかる。この課題に取り組むため、ロボットが能動的に人間に介入を求めることができる「ロボットゲート」アルゴリズム[20]、[23]～[25]への関心が高まっている。例えば、SafeDAgger [23]は、エージェントがエキスパートに制御を伝達するための信号として、エージェントがエキスパートの軌道から逸脱する可能性を決定する補助的な安全ポリシーを使用する。LazyDAgger [24]は、SafeDAggerを拡張し、エキスパート制御と自律制御の間のコンテキストスイッチの数を減らしている。ThriftyDAgger[20]と呼ばれる最近のロボットゲートアプローチは、状態が十分に新規(分布外)またはリスク(タスクの失敗につながりやすい)である場合にのみ介入することを目的としている。さらに、介入の総数をユーザーが指定した予算に制限することで、介入の負担を軽減することができる。

共変量シフト問題に取り組む2つ目の研究領域は、専門家の占有率尺度のサポートを推定し、エージェントが専門家のサポートに留まることを促す報酬を指定するアルゴリズムで構成される[17], [26], [27]。次に、報酬関数はRLを用いて最適化される。対話型ILとは異なり、これらのアルゴリズムはオンラインエキスパートへのアクセスを想定しておらず、デモンストレーションと環境とのさらなる相互作用にのみ依存している。最も一般的なアルゴリズムはIRLに基づくもので、RLエージェントが時間と共に一貫してデモに一致するように訓練することで、共変量シフトに対処する。これらの方法についての詳細な議論は、セクションIIIとIVで行う。しかし、これらの方法は、しばしば、デモから報酬関数を学習するために、敵対的学習を含む複雑で不安定な近似技法を使用する[17], [28]。本節では、RLも用いるが、報酬関数を学習する代わりに、単純な固定報酬関数を用いる、最近の代替研究ラインをレビューする。重要なアイデアは、エージェントが以下のような場合に、実証された状態に戻るよう促すことで、時間をかけて専門家の政策の支持に一貫して留まるようにインセンティブを与えることである。

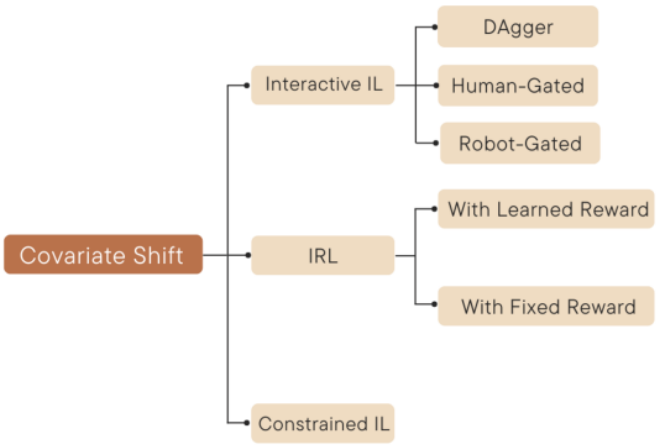

図2. 共変量シフト問題に対処する手法の分類。対話型ILはオンラインエキスパートへのアクセスを想定している。DAggerのようなアルゴリズムは、エージェントが取った各アクションに対して、専門家が修正ラベルを提供することを要求する。一方、人間ゲート法とロボットゲート法は、それぞれエキスパートとエージェントから要求された場合にのみ修正ラベルを提供する。対話型ILとは異なり、IRL手法はオンラインエキスパートにアクセスする必要がない。これらの方法は、報酬関数を最適化するための基礎となるRLアルゴリズム(デモから学習するか、固定)を必要とする。最後に、エージェントをデモでカバーされる空間の既知の領域に制約することで、制約付きILは、他の2つのカテゴリを使用して表現または解決できない共変量シフト問題に対処することを試みる。

は、専門家のサポート外の新しい状態に直面している[29]。

Wangら[26]は、主成分分析のカーネル化バージョンを用いて、エキスパートポリシーのサポートを推定する。サポート推定プロセスは、状態-アクションのペアがエキスパートポリシーのサポートに近づくにつれて増加するスコアを生成する。このスコアを用いて、固有報酬関数を構築する。

Reddyら[17]はソフトQ IL(SQIL)を提案している。SQILは、極端に疎な報酬関数を使用することで、エージェントが実演された状態のエキスパートを模倣することを促す - エキスパートの実演内部の遷移には+1の一定の報酬を、他のすべての遷移には0の一定の報酬を割り当てる。報酬関数は、エージェントが分布外の状態に遭遇した後、デモされた状態に戻ることを促す。提案モデルは単純なBCを凌駕し、限られたデモ数でも良好な性能を示す。

Brantleyら[27]は、専門家のデモンストレーションを用いて、予測値の分散をコストとして、ポリシーのアンサンブルを学習する。本質的に、アンサンブル政策は、デモンストレーションで見られなかった状態でも意見が一致しやすいので、専門家のサポート以外の分散(コスト)は高くなる。RLアルゴリズムは、教師ありBCコストと組み合わせて、このコストを最小化する。その結果、RLコストはエージェントがエキスパートの分布に戻るのを支援し、教師付きコストはエージェントがエキスパートの分布内でエキスパートを模倣することを保証する。

最後に、アルゴリズムの第3の領域は、対話的な専門家に頼ったり、RLを活用したりすることなく、エージェントをデモ機がカバーする空間の既知の領域に制約することを目的としている。これらの方法は、ヘルスケア、自律走行、産業プロセスなど、安全制約を満たさなければならない実世界のアプリケーションにとって、特に有益で実用的である[30]。

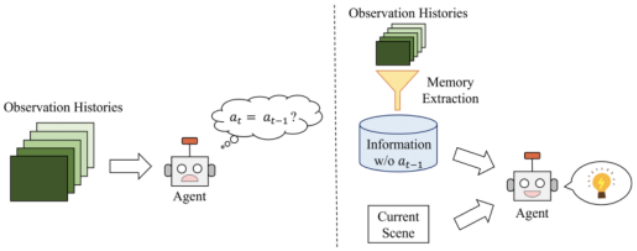

図3. 左:BCは、以前の行動を現在の行動として出力するショートカットを、以前の観察から学習するかもしれない。右:模倣のないメモリ抽出モジュール。ショートカットは、もはや履歴情報を使って利用できない[34]。

31]では、著者らは自律走行における共変量シフトの問題を克服するために、模倣損失を追加損失で補強し、悪運転を抑制することを試みている。さらに、エキスパートの軌道に対する合成摂動という形で、追加データがエージェントに提供される。摂動は衝突のような非専門家の行動にモデルをさらすので、これらの行動を避けるために追加された損失に対する重要なシグナルを提供する。

Wongら[32]は、エージェントが潜在的な故障状態にあるときを示す学習型エラー検出システムを提案している。このように、エラー検出器は、エージェントをよく知られた構成にリセットしたり、実行を終了させたりすることで、デモで以前に見た状態のみで実行するようにポリシーを制約し、潜在的に不安定な動作を防ぐことができる。[33]は、エージェントが2つのデータセット(エキスパートポリシーの状態-アクションペアの小さなデータセットと、潜在的に最適でない行動ポリシーからの状態-アクション-次の状態遷移の大きなデータセット)を提供されるオフラインIL設定を検討する。共変量シフト問題に対する彼らの解決策は、最適でないデモに対してダイナミクスモデルを学習し、データで十分にカバーされていない状態-行動空間の領域に高いペナルティを適用することからなる。

行動クローン政策で用いられる教師あり学習アプローチでは、専門家の行動の根本的な原因を特定することが難しく、「原因誤認」[35]として知られる現象につながる。学習手順において、エキスパートと環境との相互作用の因果構造は考慮されていない。したがって、クローン化された政策は、専門家の行動の真の原因から厄介な相関関係を区別できないかもしれない[35]。訓練分布とテスト分布が同じ場合、厄介な相関を無視することは、テストデータセットで保持され続けるような問題にならないかもしれない。しかし、BCでは、分布のずれのために因果関係を無視することが特に問題となる[35]。ILにおける基本的な逐次決定過程の因果構造は、この問題をさらに悪化させる。これは、因果構造-過去の行動が将来の観察に影響を与える-が、しばしばさらに複雑な厄介な相関を生み出すからである。この問題に対処するために、[35]は因果グラフからポリシーへのマッピングを学習し、最適なポリシーを見つけるために、専門家のクエリまたは環境の相互作用のいずれかのターゲット介入を使用する。Wenら[36]によるその後の研究は、”模倣問題 ”として知られる因果的混乱問題の顕著なクラスを探求している。

この問題は、専門家の行動が時間的に高い相関を持つ場合に起こる。このシナリオでは、エージェントは専門家の以前の行動をコピーすることで、不正を学習する。模倣問題に対処するために、[36]は、次の行動を予測するために必要な情報を保持しながら、既知の厄介な相関関係(前の行動)に関する情報を無視する特徴表現を学習するための敵対的アプローチを提案する。Chuangら[34]は[36]を高次元画像観測に拡張している。記憶抽出モジュールを用いて、過去の行動に関する情報を可能な限り取り除きながら、観測履歴から過去の特徴を抽出しようとする(図3)。

従来、BCは明示的なニューラルネットワークの学習を行う[37]。残念ながら、従来の明示的モデルは不連続性をモデル化するのに苦労しており、ポリシーを異なる動作の間で決定的に切り替えることができない。この問題は、連続的な活性化関数で構築されたニューラルネットワークでは、明示的なモデルでは不連続性を表現することができないために生じる。対照的に、陰的モデルは、ネットワークに連続的な層しかないにもかかわらず、鋭い不連続性を表現することができる[37]。

38]で示された暗黙のBCモデルは、観測と行動の両方を入力とし、専門家の行動では低く、専門家でない行動では高い単一の値を出力するニューラルネットワークを訓練することによって、BCをエネルギーベースのモデリング問題[39]に変える[37]。学習された暗黙のBCポリシーは、与えられた観測に対して最も低いスコアを持つアクション入力を選択する。この方法は、学習時と推論時の両方で、明示的なBCモデルよりも多くの計算を必要とする。しかし、この結果は、実世界とシミュレーションの両方において、ロボット操作タスクにおいて、従来の明示的なベースラインをしばしば上回ることができることを示している。

III. 逆強化の獲得

行動クローニングに加えて、模倣学習のもう一つの重要なアプローチはIRLである[40]。IRLは、観察されたデモンストレーションの基礎となる報酬関数を推論することを目的とする見習いエージェントを含み、これは最適に行動する専門家から来ると仮定される[41]。報酬関数が推論されると、RL [42]によって徒弟政策を訓練するように最適化される。

RLエージェントは、BCのエージェントとは異なり、環境と継続的に相互作用し、行動の結果を観察し、長期的な累積報酬を最大化するように行動を変化させることによって学習する[43], [44]。このプロセスでは、強化信号を用いて各行動の長期的な結果を学習し、エージェントがミスから回復できるようにする[27]。この能力により、IRLはBCと比較して共変量シフトの影響を受けにくい[14]。

IRLは、ロボット操作、自律航法、ゲームプレイ、自然言語処理など、様々なアプリケーションで広く利用されている[45]-[47]。それにもかかわらず、実演から学習するための効果的なIRLアルゴリズムを考案することは、主に2つの大きな理由から、困難な課題である。

第一に、IRLは計算コストが高く、リソースを大量に消費する。これは、報酬関数を正確に推定するために、

エージェントが環境と繰り返し相互作用しなければならないという事実によるところもある[19], [48]。さらに、このプロセスの性質は、特に自律走行や航空機制御のようなリスクの高いアプリケーションを扱う場合、本質的に安全でない可能性がある[49]。

さらに、典型的なIRLアプローチは、報酬推定とポリシー学習を交互に行う反復プロセスを踏襲しており、その結果、サンプル効率が悪くなる[46], [47], [50]。その結果、学習されたポリシーの安全性と精度を維持しながら、IRLアルゴリズムのサンプル効率を向上させるために、これらの問題に対処することを目的とした重要な研究が行われている。これらのアプローチの中には、報酬関数を正確に推定するために必要な相互作用の数を減らすために、人間のガイダンスを利用する方法を含むものがある[51]。

IRLの2つ目の大きな課題は、ポリシーと報酬関数の関係に固有の曖昧さがあるために生じる。具体的には,報酬関数の無限個数に関して,政策が最適であることができる[50],[52].この課題を解決するために、研究者は報酬関数に追加構造を導入する様々な方法を提案している。この曖昧さに対処することを目的としたIRL手法には、おおよそ3つのカテゴリーがある[53]。

最初のカテゴリーは最大マージン法である。最大マージン法における重要な考え方は、他の全ての政策よりもマージンをもって最適な政策をより詳細に説明する報酬関数を推論することである。これらの方法は、あるマージンを最大化する解に収束することで、議論された曖昧性問題に対処している。このカテゴリーの基礎となる方法は、Ngらの研究[50]である。彼らは、与えられた政策が最適である報酬関数を、マージンを最大化しながら線形プログラムを用いて推定する。もう一つの主要な研究は最大マージン計画(MMP)[54]であり、これは推定された方針が実証された行動に「近い」ように、特徴から報酬への重み付き線形マッピングを見つけようとするものである。[55], [56] は、関数勾配の技法のファミリーを利用することで、MMP を非線形仮説空間に構築し、拡張している。

特徴ベースの報酬関数の採用は、マージン最適化のために特徴期待値を利用する様々なアプローチを生み出した。AbbeelとNg [45]は、エキスパートのポリシーへのアクセスを仮定することなく、特徴期待損失マージンを最大化するための2つの基礎的な方法(max-marginとprojection)を提案する。他の多くのIRL手法と同様に、これらの手法にはエージェントの性能をエキスパートの品質に制限するという欠点がある。この限界に対処するために、SyedとSchapire [57]は、エキスパートよりも優れたパフォーマンスを持つポリシーを訓練することができるゲーム理論的アプローチを提案する。

IRLアルゴリズムの第二のカテゴリーは、結果として得られるポリシーのエントロピーを最大化することによって、曖昧性問題を解決することを目的とする。MaxEntIRL [47]は最大エントロピーを利用した最初のIRL手法である。Ziebartの研究[47]は、最大エントロピーパラダイムが、可能な軌道上の分布を使用することで、専門家の最適性と確率性を扱うことができることを実証している。その後の研究[58]、[59]は、パス積分法を用いてMaxEntIRLアルゴリズムを連続状態-アクション空間に拡張したものである。

完全前進マルコフ決定過程(MDP)を反復的に最適化することは、高次元の連続状態行動空間では困難になる。この複雑さを克服するために、これらの研究は実証された軌道の局所最適性を利用する。

多くの先行手法では、詳細な特徴はドメイン知識を用いて手動で抽出され、ロボットのボール・イン・カップ・ゲーム[60]のために、ボールとカップの距離のような報酬に線形結合することができる。線形表現は多くの領域では十分であるが、複雑な実世界のタスク、特に報酬値が生の感覚データから導出される場合、過度に単純化される可能性がある。Wulfmeierら[61]は、複雑で非線形な報酬関数をモデル化するためにニューラルネットワークを利用するMaxEntIRLの一般化である最大エントロピー深層IRLを提案している。また、事前に抽出した特徴量を用いる代わりに、深層アーキテクチャをさらに拡張し、畳み込み層を通して特徴量を学習する。これは学習プロセスを自動化するための重要なステップである[12]。Finら[62]による更なる研究は、[61]が報酬関数を推定するために多数のエキスパート遷移に依存しているため、[61]のサンプリング効率を改善するガイド付きコスト学習(GCL)を提案している。GCLは、(初期のIRL手法とは対照的に)政策最適化の内部ループで非線形報酬関数を学習する。これにより、複雑な制御問題に効果的に拡張することができる。この方法のさらなる利点は、報酬関数を構築するために、あらかじめ定義された特徴の代わりに、システムの生の状態を利用することであり、これは工学的負担を軽減する。

ベイズアルゴリズムはIRLアルゴリズムの第3のカテゴリーを構成する。このカテゴリのメソッドは、報酬関数の推定値を更新するための証拠として、専門家の行動を使用する。報酬関数候補の事後分布は、報酬に関する事前分布と報酬仮説の尤度から導かれる。長年にわたり、尤度に関する様々なモデルが提案されてきた。BIRL [63]は、尤度をモデル化するためのボルツマン分布に基づく、最も初期のベイズIRL手法である。報酬関数に対する事前分布として、様々な分布を用いることができる[63]。例えば、報酬が大きい二分法の問題を計画する場合、ベータ分布が適切である。報酬関数の連続空間における事後分布を解析的に求めることは非常に困難である。この問題に対処するために、[63]はマルコフ連鎖モンテカルロ法(MCMC)を用いて、事後平均の標本ベースの推定値を導出している。事後平均を計算する代わりに、[64]は最大事後(MAP)報酬関数を計算する。[64]は、事後平均は、観測された振る舞いと一致しないものであっても、報酬空間全体にわたって損失関数に積分するため、報酬推論に最も適したアプローチではないと論じている。Levineら[65]は、報酬を表現するために特徴の非線形関数を使用するベイズIRLアルゴリズムを提案する。報酬値に関するガウス過程事前分布とカーネル関数を用いて、報酬の構造を決定する。カーネルのハイパーパラメータはベイズGPの枠組みを通して学習され、報酬構造の学習につながる。

古典的なBIRLアルゴリズムは、各事後標本を生成するためにMDP全体を解く必要があるため、複雑な高次元環境では計算不可能である。この制限により、これらの手法は小さな表形式の設定を超えてスケーリングすることができない。

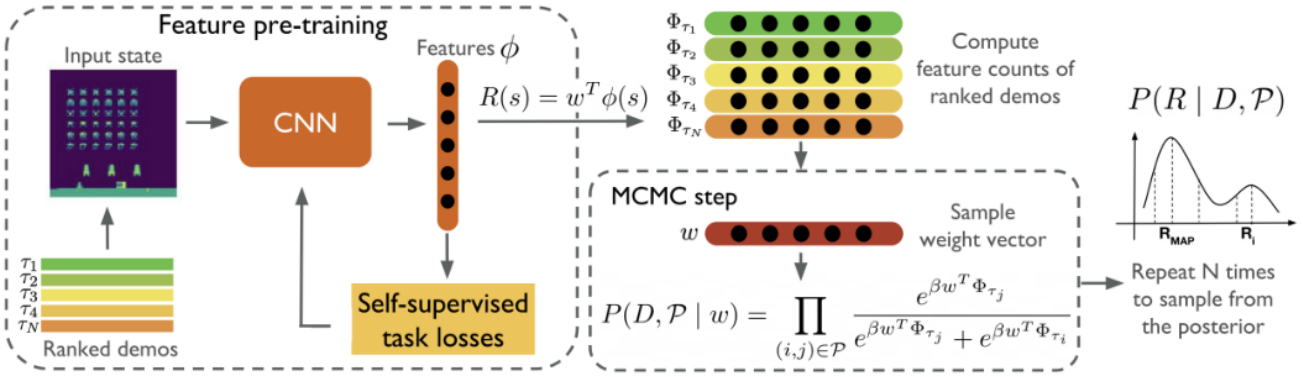

図4. 低次元の状態特徴埋め込みは、ランク付けされたデモ[66]を用いて事前に学習される。学習された特徴量の線形結合を用いて報酬関数を導出する。MCMCの提案評価では、提案(w)が与えられた場合のデモに対する選好の尤度を推定するために、ペアワイズランキング尤度が使用される。ランク付けされたデモの事前計算された埋め込みを利用することで、MCMCサンプリングは非常に効率的になる。推論中やMDPソルバーでのデータ収集は必要ない。

この制限を克服するために、[66]は、デモに対する選好ラベルを活用する代替尤度定式化を提案することで、MDPソルバーを使用せずに事後分布からサンプルを生成する(図4)。スケーラビリティの問題に対処するもう一つのアプローチは、近似変分報酬IL(AVRIL)[67]である。このアプローチは、オフライン設定において、模倣方針と報酬に対する近似的な事後分布を同時に学習する。従来のサンプリングやMAPベースの技術とは異なり、AVRILは変分推論に依存して事後分布を正確に推定する。

既存のIRLアルゴリズムのほとんどは、遷移モデル、時には専門家の方針が事前に分かっているという非現実的な仮定に依存している [47], [50], [54], [65].しかし、実世界のシナリオでは、エージェントはしばしばサンプルからエキスパートポリシーと遷移ダイナミクスを推定しなければならず、回復された報酬関数に誤差が生じる[68], [69]。彼らの分析では、[69]はこの誤差をエキスパート政策と遷移モデルの推定から構成要素に分解している。表形式の設定における)分析に基づき、学習した報酬関数を完全に既知のターゲット環境に転送することに焦点を当てた効率的なサンプリング戦略を提案する。ただし、エージェントは任意の状態や行動に対して遷移ダイナミクスを問い合わせることができると仮定する。この仮定を取り除くために、IRLのための能動的探索(AceIRL)[68]は、効率的な探索戦略の開発に焦点を当てている。この戦略は、任意のIRLアルゴリズムが報酬関数を可能な限り効果的に推論できるように、環境ダイナミクスとエキスパートポリシーの両方を探索することを目的とする。AceIRLは、これまでの観測結果を利用して、実現可能な報酬関数を捉える信頼区間を構築し、環境の最も関連性の高い領域を優先する探索方針を見つける。

IV. 敵対的模倣の獲得

IRLアルゴリズムをより大きな環境に拡張することは、専門家の行動を再現するポリシーを生成することに成功しているにもかかわらず、大きな課題となっている[62], [70], [71]。この課題は、多くのIRLアルゴリズムの計算複雑性に起因しており、多くの場合、RLは内部ループで実行される必要がある[46]。

AILは、各反復でRL部分問題を完全に解くことなく最適な方針を探索することで、IRLの計算上の課題に対する有望な解決策を提供する[46]。AILは、エージェントと敵対者(識別器)の間の2人ゲームに関与し、敵対者はエージェントの軌道と専門家の軌道を区別しようとする[72]。一方、エージェントは、専門家の軌跡に近い軌跡を生成することで、敵を欺こうとする。この敵対的なプロセスを通じて、エージェントは専門家の行動の模倣を徐々に改善し、専門家の政策に酷似した政策に収束していく。AILは、ロボット工学、自律走行、ゲームプレイなど、複数のベンチマーク環境において、既存の手法よりも統計的に有意な改善を実証している[46], [73], [74]。

IRLの限界に対処するAILの有効性は、この分野における継続的な研究に拍車をかけている。最初に注目されたAIL手法は、生成AIL(GAIL)[46]として知られている。GAILでは、報酬関数はエージェントが専門家の行動を模倣する能力を測定する。そのために、GAILはエキスパートの行動とエージェントが生成した軌跡を区別するように訓練された識別器ネットワークを利用する。報酬信号は、エージェントによって生成された軌跡か、エキスパートによって生成された軌跡かを判断することがいかに困難であるかを反映し、識別器の混乱から導かれる。この報酬信号を最大化することで、エージェントは専門家の行動に近い軌道を生成するインセンティブを持つ。長年にわたり、識別器の損失関数の変更[76]や、オンポリシーからオフポリシーへの切り替え[77]など、サンプル効率、スケーラビリティ、ロバスト性を向上させるために、オリジナルのアルゴリズムに多くの改良が提案されてきた[75]。

AILでは、エージェントが専門家の軌跡に近い軌跡を生成できるようにすることが目的である。これは、両者の類似性を定量化するために距離尺度を使用することを含む。様々なAIL手法は、エージェントが遭遇する状態や行動に関する分布を専門家の分布と一致させるために、様々な類似性尺度を採用している[29]。例えば、GAILはシャノン・ジェンセン発散を利用し、AIRL [76]のようないくつかの手法はカルバック・ライブラー発散を利用する。しかし、Arjovskyらによる最近の研究[78]では、f-ダイバージェンスを

Wasserstein距離をその双対定式化によって学習の安定性を向上させることができ、この手法はいくつかのAIL手法で実装されている[77], [79]。これらの開発を踏まえると、新しい類似性尺度を探求することは、新しいAIL手法を発見する可能性を秘めている。

ほとんどのAIL手法は、GAN(Generative adversarial networks)[80]と同様に、エキスパートとエージェントの状態-行動分布間の距離を最小化する一方で、識別器の混乱に由来する報酬信号を最大化するために、最小-最大最適化アプローチを使用する。しかし、このアプローチは、勾配の消失や収束の失敗などの問題があるため、学習が困難である可能性がある[28]。これらの課題を克服するために、原始ワッセルシュタインIL(PWIL)[29]のような、原始-双対アプローチによってワッセルシュタイン距離を近似する方法が開発されている。

V. Imitation from Observation

ILにおける一般的なパラダイムは、学習者が専門家によって示された状態と行動の両方にアクセスできることを前提としている[81]。しかし、そのためには、IL目的のために明示的にデータを収集する必要があることが多い[81]。例えばロボット工学では、エキスパートはロボットを遠隔操作するか、手動で関節を動かさなければならない(運動学的学習)[82]。いずれの場合も、かなりのオペレータの専門知識が必要であり、有用なデモンストレーションは人工的な条件下で記録されたものに限定される。これらの制限要因は、専門家の行動が未知であるIfO [83]における最近の取り組みの動機となっている。これまでの方法とは対照的に、観察からの模倣は専門家から学ぶためのより自然な方法であり、人間や動物が模倣一般にどのようにアプローチするかに同調する。人間にとって、低レベルの行動(例えば筋肉の命令)に気づかないまま、他の人間を観察することで新しい行動を学習することは一般的である。人間は、動画をオンラインで見ることで、織る、泳ぐ、ゲームをするなど、幅広いタスクを学習する。体型、センシングモダリティ、タイミングには大きなギャップがあるかもしれないが、オンラインデモンストレーションから得られた知識を適用する驚くべき能力を示している[9]。

エージェントが行動情報なしにデモから学習できるようにすることで、インターネット上のビデオなど、これまで適用できなかった多くのリソースを学習に利用できるようになる[84]。さらに、アクションが未知であったり、マッチングできなかったりする、異なる実施形態を持つエージェントから学習する可能性を開く。ILのための状態のみのデモの使用は新しいものではありません[85]。しかし、最近の深層学習と視覚認識の発展[86]は、特に生の視覚観測を扱う場合、問題にアプローチするためのより強力なツールを研究者に装備している[48]。

Liuら[83]は、文脈を考慮した翻訳を用いて、生映像から模倣者方針を学習する観察からの模倣法を提案している。彼らのアルゴリズムは、エキスパートのコンテキスト(例えば、三人称視点)からエージェントのコンテキスト(例えば、一人称視点)へデモを変換するコンテキスト変換モデルを利用する。次に、このモデルを用いて、ロボットの文脈における専門家の行動を予測する(図5)。

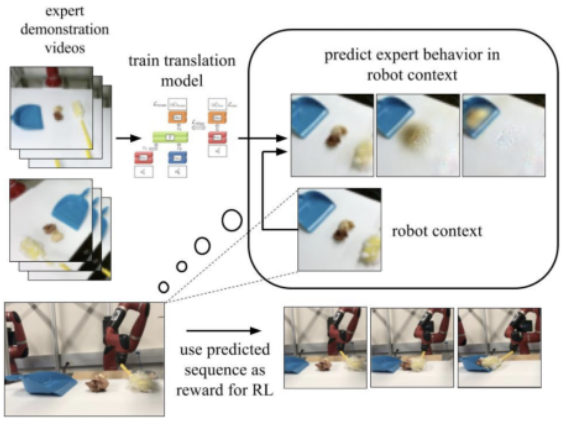

図5. 文脈変換モデルは、専門家のデモンストレーションのいくつかのビデオで学習される[83]。ロボットは学習プロセス中に実行しなければならないタスクのコンテキストを観察する。そして、モデルは、ロボットの文脈で専門家が何をするかを決定する。

予測されたオブザベーションを使用して、報酬関数が定義され、これは、入力オブザベーションからエンコードされたエキスパートの変換された特徴から逸脱するためのペナルティと、変換されたオブザベーションとは異なるオブザベーションに遭遇するためのペナルティで構成される。RLは、導出された報酬関数を最適化するために使用される。この方法の適用を制限する2つの欠点がある。まず、異なるコンテキストからのデモは時間的に整列していると仮定するが、これは現実世界ではほとんど当てはまらない[87]。第二に、翻訳モデルの学習には、多数のデモが必要である[83]。さらに、実施形態の違い[83]のような体系的なドメインシフトに対応できないという限界もある。

Sermanetら[88]は、異なる視点や実施形態に対して不変な、時間対比ネットワーク(TCN)を用いた自己教師付き表現学習法を紹介している。TCNは、カメラの角度など、文脈の違いに不変な特徴を抽出するために、各ビデオフレームの埋め込みを学習するニューラルネットワークを学習する。三重項損失関数を用いることで、異なるモダリティ(すなわち、視点)を持つ同じ時間に発生する2つのフレームは、埋め込み空間において、離れた時間ステップのフレームは、視覚的に類似したフレームを持つが、より近くに押し出される(図6)。報酬関数を構築するために、デモの埋め込みとエージェントのカメラ画像の埋め込みとの間のユークリッド距離を計算する。模倣方針を学習するための報酬関数を最適化するために、RL技術が使用される。この手法の限界は、トレーニングのためにマルチビューポイントビデオを必要とすることであり、これは(例えばインターネット上で)容易に利用できない。

BC from observation (BCO) [89]は、行動クローニングアプローチをとることで、先行手法のRLアルゴリズムの学習に必要な、実証後の環境相互作用の量を最小化することを目的としている。BCOはまず、最初にランダムな方針に従い、環境と相互作用し、データを収集するエージェントに、逆ダイナミクスモデルを学習させる[81]。その後、モデルを用いて、専門家のデモンストレーションの欠落した行動を推論する。次に、BCアルゴリズムを用いて、推論された行動に状態をマッピングし、通常のIL問題として解く。

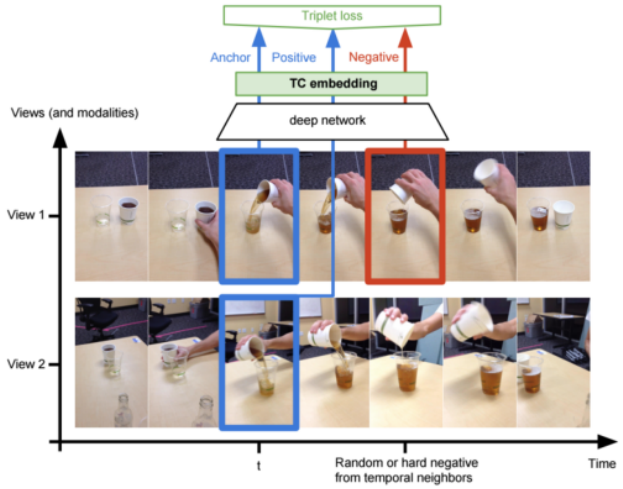

図6. 埋め込み空間は、異なる視点から撮影された画像が互いに近接するように、異なる視点から撮影された画像が遠距離になるように促す[88]。

このアプローチを用いると、特に高次元問題において、ダイナミクスモデルをオンラインで学習するために大量のデータを収集する必要がある。

その後の[81]の研究では、オフラインで潜在的なフォワードダイナミクスモデルを学習することで、BCOで必要とされる環境相互作用の数を減らすことを試みる[48]。この研究の基礎となる仮定は、未知ではあるが、予測可能な原因が、観測された状態遷移のクラスを記述するというものである。目標は、エージェントがこれらの潜在的な原因を予測し、模倣できるようにすることである[81]。これを実現するために、潜在的な政策が学習され、潜在的な行動が観測された状態で取られる確率を推定する。そして、エージェントが取り得る実世界の行動と、モデルによって識別された潜在的な行動との間のマッピングを学習するために、環境との限られた数の相互作用を使用する。

Generative adversarial immitation from observation (GAIfO) [84]は、エキスパートとエージェントの状態遷移分布をマッチングさせることで、GAILの目的をIfOに適応させる。敵対的アプローチを採用することで、この方法は以前のアプローチ[81], [89]で遭遇した共変量シフトの問題を克服することができる。また、従来のアプローチとは異なり、時間的にずれたデモを扱うことができる。このアプローチを用いると、エキスパートとエージェントが同じ環境で、同じダイナミクスの下で動作するときに、最も成功する。しかし、エキスパートの状態遷移はエージェントの環境では実現不可能である可能性があるため、ダイナミクスが異なる場合、状態遷移分布を一致させることはより困難になる[90]。

Jaegleら[91]は、尤度ベースの生成モデルを用いた観測からの非逆説的IRLアプローチを紹介している。この方法では、条件付き状態遷移確率をエキスパートと学習者の間でマッチングさせる。著者らの知見によると、条件付き状態遷移確率をマッチングさせる彼らのアプローチは、状態-次状態結合確率をマッチングさせるGAIfOのような敵対的アプローチよりも、エキスパートと学習者の設定間の無関係な差異に焦点を当てる傾向がある。特に、条件付き状態確率は、デモには存在しないが、正しい遷移を導く特徴に誤ったペナルティを与えにくいと主張する。

Raychaudhuriら[87]は、対になっていない、整列されていないデモから、ソースとターゲットのドメインにまたがる状態マップを学習するフレームワークを提案している。このアプローチは、具現化、視点、ダイナミクスのミスマッチに対処する。ドメイン変換中にMDPダイナミクスを保持するために、ローカルアライメントとグローバルアライメントが実行される。ローカルアライメントでは、真の軌道と転送された軌道の状態遷移分布の乖離を最小化するために、敵対的学習が用いられる。一方、学習された時間位置関数は、2つのドメイン間で一貫した時間位置に状態が配置されることを保証するために、グローバルアライメントを強制するために使用される。最後に、転送されたデモのセットが与えられると、BCOは最終的なポリシーを学習するために使用される。83]と同様に、この方法は代理タスク、すなわち、両方のドメインからの専門家のデモンストレーションに依存しており、その適用が制限される。

Aytarら[9]は、YouTubeの動画を見て、Atari環境に明示的に触れることなく、ハードな探索Atariゲームをプレイすることを学習するための新しい自己教師付きフレームワークを提示している(図7)。YouTube動画からの学習は、ドメイン固有のバリエーション(例えば、色や解像度)の存在や、フレームごとのアライメントの欠如により、いくつかの課題を提起する。これらの課題に対処するために、彼らはまず、視覚と音声の両方で構築された自己教師付き分類タスクを使用し、複数のソースからの整列されていないビデオを共通の表現にマッピングする。次に、この表現にYouTubeの動画を1つ埋め込み、その埋め込みに沿ってチェックポイントのシーケンスを配置する。これは、エージェントが人間のゲームプレイを模倣することを促す報酬関数を作成するために使用される。政策学習中、エージェントはこれらのチェックポイントに到達したときのみ報酬を得る。

Brownら[97]は、最適でないランク付けされたデモンストレーションからエキスパートの意図を外挿するための、観察からのIRL技術を導入している。本研究では、高次元タスクにおいて、専門家の意図を推論することで、最適でない専門家に対する性能を向上させることを目的とする。彼らは、より上位の軌道により大きなトータルリターンが割り当てられるように、状態ベースの報酬関数を学習する。このようにランキングを利用して報酬関数を構築することで、ランキングと相関のある特徴を特定することができ、デモンストレータよりも優れた性能を発揮できる可能性がある。学習された報酬関数が与えられると、RLはポリシーを最適化するために使用される。

YouTubeの大量のナビゲーションデータを利用し、[98]はスケーラブルな運転を学習するためのフレームワークを提案している。まず、単眼画像を鳥瞰図(BEV)にマッピングするために、小さなラベル付きデータセットでモデルを学習し、YouTube動画の制約のない性質(例えば、視点やカメラパラメータ)からの学習を容易にする。公開されている運転データセットには行動ラベルが含まれていることが多いので、この仮定は妥当である。この学習済みモデルは、大規模なラベルなしデータセットに対して擬似ラベルを生成するために使用される。最後に、擬似ラベル付きデータセットで一般化されたポリシーを学習し、小さなラベル付きデータセットのクリーンラベルで微調整を行う。

VI. 課題と限界

A. 不完全なデモンストレーション

IL法では、専門家であるデモンストレーターが行う最適なデモンストレーションを想定している。

TABLE I

S模倣学習に関する既存の研究のまとめ

Ref	Datasets	Inputs	Learning Type	Online/ Offline	Online Expert	Application
[93]	Sim	State, Image	Interactive IL	Online	Yes	ロボット運動、係り受け解析
[17]	Sim	State, Image	Regularized BC	Online	No	カーレース、アタリゲーム、ロコモーションコントロールタスク
[14]	Sim	State, Image	BC\DAgger	Online	Yes	ゲーム、手書き文字認識
[21]	Sim, Real	State	BC\HG-DAgger	Online	Yes	Autonomous Driving
[22]	Sim	State	BC\ ヒトゲート型インタラクティブIL	Online	Yes	Robotic Manipulation
[19]	Sim	State	Human-in-the-loop RL	Online	Yes	Autonomous Driving
[23]	Sim	Image	BC\SafeDAgger	Online	Yes	Autonomous Driving
[24]	Sim, Real	State, Image	BC\LazyDagger	Online	Yes	Robotic Locomotion, Fabric Manipulation
[25]	Sim	State	BC\EnsembleDagger	Online	Yes	Inverted Pendulum, Locomotion
[20]	Sim, Real	State, Image	BC\ThriftyDAgger	Online	Yes	Peg Insertion, Cable Routing
[26]	Sim	State	IL via Expert Support Estimation	Online	No	Robotic Locomotion, Autonomous Driving
[27]	Sim	State, Image	IL via Expert Support Estimation	Online	No	Atari Games, Continuous Control Tasks
[31]	Sim, Real	State	IL via 専門家によるサポートの推定	Online	No	Autonomous Driving
[32]	Sim	State	BC	Online	No	Robotic Manipulation
[33]	Sim	State	オフライン模倣学習	Offline	No	Robotic Locomotion
[35]	Sim	State	因果グラフパラメータ化政策学習	Online	Yes	Atari Games, Robotic Locomotion
[36]	Sim	State	BC	Offline	No	Robotic Locomotion
[34]	Sim	State, Image	BC	Offline	No	Robotic Locomotion, Autonomous Driving
[38]	Sim, Real	State, Image	Implicit BC	Offline	No	Robotic Manipulation
[94]	Sim, Real	State	Maximum Entropy IRL	Online	No	Path Planning, Gridworld
[45]	Sim	State	Feature Based IRL	Online	No	自律走行、グリッドワールド
[47]	Real	State	最大エントロピーIRL	Online	No	運転行動の予測、ルート推薦
[54]	Real	State	最大マージンIRL	Online	No	Route Planning
[55]	Sim, Real	State, Image	最大マージンIRLMMBOOST	Online	No	経路計画、脚式ロコモーション、運転障害物検知/回避
[56]	Sim, Real	State, Image	最大マージンIRL	Online	No	足踏み予測、把持予測、ナビゲーションタスク
[57]	Sim	State	最大マージンIRL	Online	No	Car Driving Game
[58]	Sim	State	Maximum Entropy IRL	Online	No	2-D Point Mass Control System
[59]	Real	State	Maximum Entropy IRL	Online	No	Robotic Manipulation
[60]	Sim	State	Relative Entropy IRL	Online	No	Car Racing, Gridworld, Game
[61]	Sim	State	最大エントロピー深層IRL	Online	No	Objectworld, Binaryworld
[62]	Sim, Real	State	Maximum Entropy IRL	Online	No	Robotic Manipulation
[63]	Sim	State	Bayesian IRL	Online	No	Random Generated MDPs
[64]	Sim	State	Bayesian IRL\ MAP Inference	Online	No	グリッドワールド、簡易カーレース

TABLE I
S模倣学習に関する既存研究の概要(続き)

[65]	Sim	State	Nonlinear Bayesian IRL	Online	No	オブジェクトワールド、高速道路走行
[66]	Sim	Image	Bayesian IRL	Online	No	Atari Games
[67]	Sim, Real	State	Bayesian IRL	Offline	No	医療情報データセット、物理ベース制御
[69]	Sim	Image	IRL	Online	Yes	グリッドワールド、ランダム生成MDP、チェーンMDP
[68]	Sim	State	IRL\Active Exploration	Online	No	ランダムMDP、グリッドワールド チェーンMDP、ダブルチェーン
[46]	Sim	State	Generative Adversarial IL	Online	No	物理ベース制御、ロボット運動
[76]	Sim	State	Adversarial IRL	Online	No	ランダム生成MDP Continuous Control Tasks
[77]	Sim	State	Adversarial IRL	Online	No	ロボットの運動と操作
[95]	Sim	State	Bayes-GAIL	Online	No	Robotic Locomotion
[96]	Sim	State	IRL	Online	No	Robotic Locomotion
[29]	Sim	State, Image	Adversarial IRL	Online	No	ロボットのロコモーションと手の操作
[79]	Sim	State, Image	Generative Adversarial IL\ InfoGAIL	Online	No	Synthetic 2D Example, Autonomous Highway Driving
[83]	Sim, Real	Image	Imitation from Observation	Online	No	Robotic Manipulation
[88]	Sim, Real	Image	IfO\Self-supervised Learning	Online	No	ロボット操作、人間の姿勢模倣
[89]	Sim	State	BC from Observation	Online	No	物理ベース制御、ロボット運動
[81]	Sim	State, Image	観察からの模倣	Online	No	物理ベース制御、コインランゲーム
[84]	Sim	State, Image	観測からの生成的逆数的模倣	Online	No	物理ベース制御、ロボット運動
[90]	Sim	State	観察からの模倣	Online	No	Robotic Locomotion
[91]	Sim	State	IRL from Observations	Online	No	Robotic Locomotion
[87]	Sim	State	観測から得られたクロスドメインIL	Offline	No	物理ベース制御、ロボット運動
[9]	Sim	Image	観察からの模倣	Online	No	Atari Games
[97]	Sim	State, Image	IRL from Observations	Online	No	Atari Games, Robotic Locomotion
[98]	Sim, Real	Image	観測値からの条件付きIL	Offline	No	Autonomous Driving
[99]	Sim	State	Importance Weighting IL, GAIL	Online	No	Robotic Locomotion
[100]	Sim	State	ノイズの多いデモからのBC	Offline	No	Robotic Locomotion
[101]	Sim	State, Image	Weighted GAIL	Online	No	Atari Games, Robotic Locomotion
[102]	Sim	State	Weighted BC	Offline	No	Robotic Locomotion
[103]	Sim	State	BC	Offline	No	ミニグリッド環境、ロボット操作 Chess Game-endings
[104]	Sim	State	目的を修正したGAIL	Online	No	Robotic Locomotion
[105]	Sim	State	GAIL	Online	No	Physics-based Control
[106]	Sim, Real	Image	Cross-embodiment IRL	Online	No	Robotic Manipulation
[107]	Sim	State	IL via Optimal Transport	Online	No	物理ベース制御、ロボット運動、2D迷路ナビゲーション
[108]	Sim	State	BC	Offline	No	ロボット操作、物理ベース制御

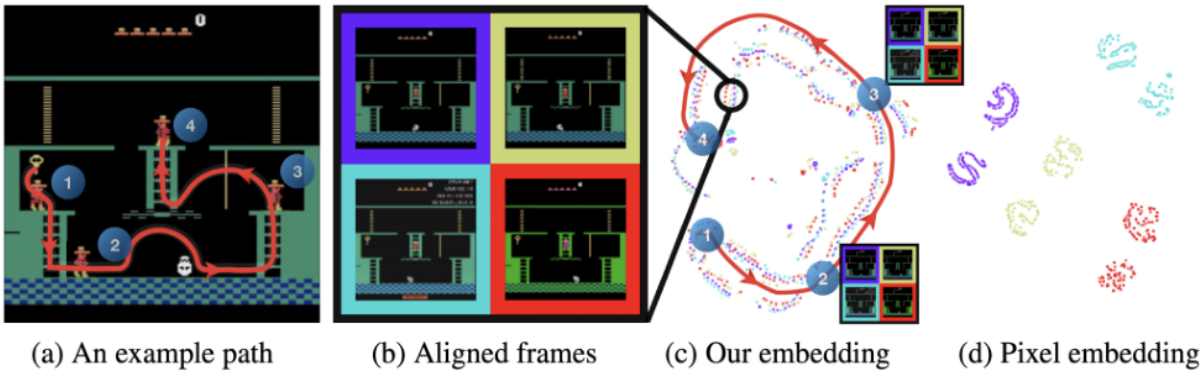

図7. (a)の経路について、提案する埋め込みを用いた軌跡のt-SNE投影[92](c)と生の画素(d)を示す[9]。bでは、MONTEZUMAのREVENGEのフレーム例を、アーケード学習環境と3つのYouTube動画という4つの異なるドメインを使って比較している。埋め込み空間から、4つの軌跡がうまく整合していることがわかる。

[2]. しかし、この仮定は、様々なケースでデモンストレーションから学習する場合、制約が多すぎる[2]。まず、人間の専門家から高品質なデモを大量に得ることは困難である[109], [110]。多くの実世界のタスクでは、必要な時間と労力の量から、これは人間にとって不可能であろう。さらに、人間は注意散漫の存在や環境の限定的な観察可能性など、様々な理由でミスを犯しやすい [99], [100]。第二に、ロバストで効果的なILポリシーを学習するために、クラウドソーシングデータセットの規模と多様性を活用する必要がある[111]。しかし、クラウドソーシングされたデータセットは、様々なレベルの専門知識を持つユーザーから収集されるため、必然的に幅広い行動最適性を持つことになる。

不完全なデモンストレーションに対する素朴な解決策は、最適でないデモンストレーションを捨てることである。しかし、このスクリーニング・プロセスは、多大な人的労力を必要とするため、しばしば非現実的である[100]。そのため、不完全なデモンストレーションから学習できる手法の開発に関心が高まっている。

Wuら[99]は、信頼度スコア付きデータとラベルなしデータの両方を利用することで、不完全なデモンストレーションに対処する2つの一般的なアプローチを提示している:2段階の重要度重み付けIL(2IWIL)と不完全なデモンストレーションと信頼度を持つ生成的敵対的IL(IC-GAIL)。どちらのアプローチも、デモの一部が信頼度スコア(すなわち、与えられた軌跡が最適である確率)でアノテーションされていることを前提としている。2IWILは、まず半教師付き分類器を用いてラベルのないデモの信頼度スコアを生成し、次に再重み付け分布を用いた標準的なGAILを実行する2段階のアプローチである[102]。IC-GAILは、2つのステップにおける誤差の蓄積を避けるために、分類器の学習を放棄し、ラベルのないデモで占有率測定マッチングを実行する。

佐々木ら[100]は、非最適なデモに関連するスクリーニングや注釈を一切用いずに、ノイズの多い専門家から得られたノイズの多いデモから学習するオフラインBCアルゴリズムを提案している。重要なアイデアは、学習されたポリシーを活用して、重み付きBCの次の反復でサンプルを再重み付けすることである。ノイズの多いエキスパートの行動分布は、最適なエキスパートの行動分布と最適でないエキスパートの行動分布の2つの分布の重み付き混合であると仮定する。目標は、ノイズの多いエキスパート行動分布モードが最適なエキスパート行動分布モードに近づくように重みを変更することである。これは、古いポリシー(すなわち、前の反復で最適化されたポリシー)を、重み付きBC目的におけるアクションサンプルの重みとして再利用することによって達成される。しかし、このアプローチは、最適なデモンストレーションがデータの大部分を占める場合にのみ、最適なポリシーに収束する。

Wangら[101]は、オラクルからの補助情報を必要とせずに、GAILの不完全なデモンストレーションに重み付けする方法を研究している。学習用の各デモンストレーションの品質と重要性を評価するために、自動重み予測法を提案する。GAILの識別器とエージェントポリシーの両方を用いて、重みを正確に推定できることを実証している。学習手順では、まず各デモンストレーションの重みを決定するために重み推定を行う。重み付きGAILを用いて、エージェントポリシーは重み付きデモンストレーションで学習される。この2つの手順は交互に作用し、全体として最適化される。

Kimら[102]は、未知の最適性レベルを持つ補足的な不完全実証を用いることで、十分な専門家による実証の欠如によって引き起こされる分布シフト問題を克服することを目的としている。彼らは、エージェント分布と専門家分布と不完全分布の混合との間のKLダイバージェンスによって、ILの分布マッチング目的を正則化する。この正則化された目的語の最適な状態-行動分布は、デュアルプログラム技法を用いて得られる[112]。最適な状態行動分布が与えられると、重み付きBCを実行することでエキスパートポリシーが抽出される。

Beliaevら[103]は、デモンストレーターの専門知識(ILEED)を推定することでILを導入している。実証者の身元に関する情報を活用し、教師なし方式で彼らの専門知識レベルを推論する。各デモンストレーターには、州に依存した専門知識値が割り当てられ、これは、どのデモンストレーターが特定の州でより良いパフォーマンスを発揮し、異なる州でそれぞれの強みを組み合わせることができるかを示している。ILEEDは、学習された政策と専門知識レベルに対する共同モデルを開発し、最適化する。その結果、モデルは各デモンストレーターの最適な行動から学習し、最適でない行動をフィルタリングすることができる。専門知識レベルは、状態埋め込みとデモンストレータ埋め込みの2つの埋め込みの内積でモデル化される。

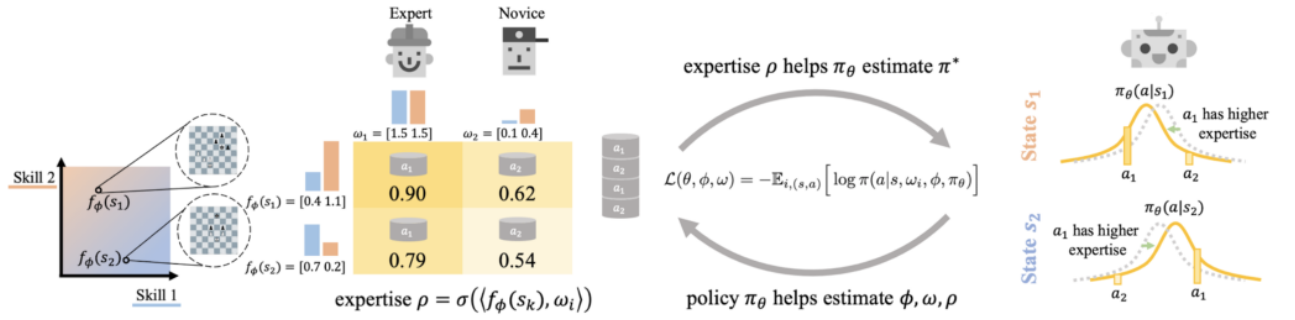

図8. デモ参加者の専門知識を推定することによるIL[103]。左:状態に関連するスキルは、状態埋め込みによって符号化される。中: ある状態におけるデモンストレーターの専門知識ρは、状態埋め込みとデモンストレーター埋め込みによって決定される。右図: 専門知識レベルを利用することで、学習されたポリシーを改善し、状態/デモの埋め込みと専門知識レベルの推定を改善する。

埋め込みベクトルの各次元は潜在的なスキルに対応し、状態埋め込みはそのスキルがその状態で正しく行動することにどれだけ関連するかの重み付けを持ち、デモンストレータ埋め込みはそのスキルにどれだけ熟練しているかを表す(図8)。

B. ドメインの不一致

先行研究の大半は、エキスパートとエージェントが同じ状態空間と行動空間の下で動作することを前提としている[107]。この仮定により、エキスパートとエージェントの行動間の一対一の対応関係を手動で指定することが容易になる。しかし、この場合、これらのアルゴリズムの適用は、専門家のデモンストレーションがエージェントドメインから来る単純なシナリオに限定される。近年、より緩やかで現実的な仮定の下で、ILへの関心が高まっている:エキスパートドメインにおけるタスクのデモンストレーションが与えられたとき、エージェントが自身のドメインでタスクを最適に実行することを学習するように訓練する[107]。この緩和された設定は、ドメイン内の専門家によるデモンストレーションの必要性を取り除き、ILの効率とスケーラビリティを向上させることで、デモンストレーションの収集を容易にする。ダイナミクス[90]、[104]、視点[88]、[105]、および具現化の不一致[106]、[107]、[113]。

IL研究における異なるドメイン間の知識の伝達には、しばしば状態行動空間間のマッピングを学習することが含まれる。最近の研究[83]、[88]は、ドメイン不変な特徴空間への状態マッピングやエンコーディングを学習するために、両方のドメイン(エキスパートとエージェント)からのペアと時間整合されたデモンストレーションを利用する。これに続いて、与えられたタスクの最終的なポリシーを学習するためにRLステップを実行する。これらの研究は、ペアとなるデモの利用可能性が限られていることと、RL手順のコストが高いことから、その適用には限界がある[108]。これらの制限を克服するために、Kimら[108]は、オンラインエキスパートにアクセスできる一方で、対になっていないデモと整列していないデモから状態マップと行動マップを学習する一般的なフレームワークを提案する。さらに、ゼロショット模倣を行うためにアクションマップを活用することで、高価なRLステップの必要性を排除している。[87]の研究は、[108]を模倣形式の観察設定に拡張し、また[108]のオンラインエキスパートの必要性を排除している。これらの方法はすべてプロキシタスクに依存しており、実世界のシナリオでの適用には限界がある。

Stadieら[105]は、異なる視点からのデータの区別に識別器を使用し、代理タスクなしでドメインの混乱を最大化する、視点にとらわれない模倣のための敵対的フレームワークを提案する。Zakkaら[106]は、きめ細かな構造的詳細のマッチングではなく、タスクの進捗を模倣することに焦点を当てたゴール駆動型アプローチを採用している。これらのアプローチのいくつかは、前のセクションですでに説明したとおりである。以下は、最新の手法のいくつかについて、詳細な考察を行うものである。Chaeら[104]は、摂動された環境ダイナミクスの下で良好なパフォーマンスを発揮できるポリシーを学習するためのフレームワークを提供している。目的は、環境ダイナミクスの連続体からわずかなサンプルを用いて、連続的なダイナミクスの変動に対してロバストなポリシーを訓練することである。サンプリングされた環境は、デモ収集フェーズと政策対話フェーズの両方で使用される(図9)。そして、この問題は、複数のエキスパートポリシーとエージェントポリシーの間のJensenShannon発散の加重平均の最小化として定式化される。Cross-embodiment IRL (XIRL) [106]は、異なるエージェントが異なる方法で同じタスクを実行する動画から、エージェント不変のタスク定義を抽出しようとするものである。XIRLは、時間サイクル一貫性(TCC)を用いて視覚的埋め込みを学習し、様々な長さのビデオにおける重要な瞬間を特定し、タスクの進行を符号化するためにそれらをクラスタリングする。具現化不変の報酬関数を学習するために、XIRLはTCC埋め込み空間における単一のゴール状態からの距離を用いる。この方法は、エキスパートと学習者の間でビデオフレームを手動でペアリングする必要がないため、(スキルレベルに関係なく)任意の数の実施形態またはエキスパートに適用することができる。Fickingerら[107]は、明示的なクロスドメイン潜在空間[106]に頼ったり、あらゆる形態の代理タスク[83]、[87]、[88]に頼ることなく、異なる実施形態を持つ模倣エージェントを訓練するために、専門家のデモンストレーションをどのように使用できるかを検証している。その代わりに、エキスパートとエージェントの状態-行動占有率間のグロモフ・ワッサーシュタイン距離を用いて、2つのドメイン間の距離尺度を保存する等角変換を求める。エキスパート領域とエージェント領域からの軌跡が与えられたとき、

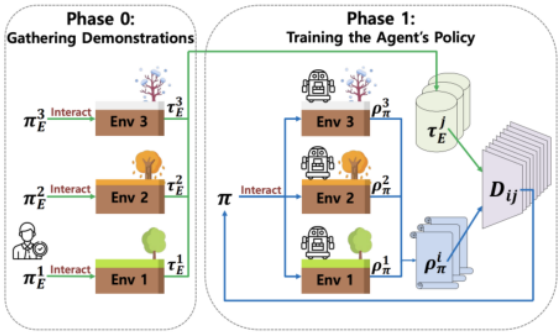

図9. 環境ダイナミクスの変動に対するIL [104]. 青はポリシーサンプルの流れを表し、緑はエキスパートデモの流れを表す。

エージェント領域での状態からその近傍までの距離がエキスパート領域でどの程度保存されるかに基づいて擬似報酬が計算される。これらの擬似報酬を用いて、RLアルゴリズムを用いてポリシーを最適化する。

VII. 機会および今後の課題

このサーベイ論文は、ILのアルゴリズム、分類、開発、および課題を探求しながら、ILの分野の包括的な概要を提供する。本稿ではまず、ILアルゴリズムの分類を提示し、2つの一般的な学習アプローチ、すなわちBCとIRLを特定し、それらの相対的な利点と限界について議論する。さらに、本論文は、敵対的訓練をILに統合することの利点を強調し、AIL分野における現在の進歩を評価する。また、状態のみのデモから学習することを目的とした、If0と呼ばれる新しい手法も紹介する。

様々なILアルゴリズムの検討を通じて、我々はその長所と限界について貴重な洞察を得て、今後の研究のための重要な課題と機会のいくつかを特定した。ILアプローチのあらゆるカテゴリーにわたる重要な課題の1つは、多様で大規模なデモンストレーションを収集する必要性であり、これは実世界で適用可能な一般化可能なポリシーを訓練するために極めて重要である[111]。しかし、オンラインビデオのような容易に入手可能なデモリソースは、デモ参加者の間で専門知識のレベルが異なるなど、さらなる困難をもたらすため、これは課題を提起する。

IL研究におけるもう一つの課題は、エージェントがダイナミクス、視点、具現化において異なる領域を横断して学習できるような手法を開発することである。エージェントに専門家から効果的に学び、IL研究からの洞察を実世界のシナリオに適用することを教えるには、これらの課題を克服することが不可欠である。したがって、今後の研究では、不完全なデモンストレーションから学習し、有用な情報を抽出し、クロスドメイン学習を可能にするアルゴリズムの開発に焦点を当てるべきである。このような課題にもかかわらず、ILの分野は将来の研究のためのエキサイティングな機会を提示している。AIの分野が進化し成熟し続けるにつれて、ILはエージェントがデモから学習し、新しいタスクや環境に適応し、最終的にはより高度なレベルの知能を達成できるようにする上で重要な役割を果たすと我々は考えており、AIの実世界での応用への道を開くものである。

ACKNOWLEDGMENTS

本研究の一部は、オーストラリア研究評議会の発見プロジェクト資金スキーム(プロジェクトDP190102181およびDP210101465)の支援を受けた。

REFERENCES

[1] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[3] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[4] M. Deng, Z. Li, Y. Kang, C. P. Chen, and X. Chu, "A learning-based hierarchical control scheme for an exoskeleton robot in human–robot cooperative manipulation," *IEEE transactions on cybernetics*, vol. 50, no. 1, pp. 112–125, 2018.

[5] Z. Zhu and H. Hu, "Robot learning from demonstration in robotic assembly: A survey," *Robotics*, vol. 7, no. 2, p. 17, 2018.

[6] J. Van Den Berg, S. Miller, D. Duckworth, H. Hu, A. Wan, X.-Y. Fu, K. Goldberg, and P. Abbeel, "Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 2074–2081.

[7] L. Le Mero, D. Yi, M. Dianati, and A. Mouzakitis, "A survey on imitation learning techniques for end-to-end autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[8] P. M. Kebria, A. Khosravi, S. M. Salaken, and S. Nahavandi, "Deep imitation learning for autonomous vehicles based on convolutional neural networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 1, pp. 82–95, 2020.

[9] Y. Aytar, T. Pfaff, D. Budden, T. Paine, Z. Wang, and N. De Freitas, "Playing hard exploration games by watching youtube," *Advances in neural information processing systems*, vol. 31, 2018.

[10] U. E. Ogenyi, J. Liu, C. Yang, Z. Ju, and H. Liu, "Physical human–robot collaboration: Robotic systems, learning methods, collaborative strategies, sensors, and actuators," *IEEE transactions on cybernetics*, vol. 51, no. 4, pp. 1888–1901, 2019.

[11] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," *IEEE transactions on cybernetics*, vol. 50, no. 8, pp. 3668–3681, 2019.

[12] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[13] D. A. Pomerleau, "Efficient training of artificial neural networks for autonomous navigation," *Neural computation*, vol. 3, no. 1, pp. 88–97, 1991.

[14] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[15] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[16] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[17] S. Reddy, A. D. Dragan, and S. Levine, "Sqil: Imitation learning via reinforcement learning with sparse rewards," in *International Conference on Learning Representations*, 2020.

[18] J. Roche, V. De-Silva, and A. Kondoz, "A multimodal perception-driven self evolving autonomous ground vehicle," *IEEE Transactions on Cybernetics*, 2021.

[19] Q. Li, Z. Peng, and B. Zhou, "Efficient learning of safe driving policy via human-ai copilot optimization," in *International Conference on Learning Representations*, 2022.

[20] R. Hoque, A. Balakrishna, E. Novoseller, A. Wilcox, D. S. Brown, and K. Goldberg, "Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning," in *Proceedings of the 5th Conference on Robot Learning*, vol. 164. PMLR, 2021, pp. 598–608.

[21] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "Hg-dagger: Interactive imitation learning with human experts," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.

[22] A. Mandlekar, D. Xu, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese, "Human-in-the-loop imitation learning using remote teleoperation," *arXiv preprint arXiv:2012.06733*, 2020.

[23] J. Zhang and K. Cho, "Query-efficient imitation learning for end-to-end simulated driving," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 2891–2897.

[24] R. Hoque, A. Balakrishna, C. Putterman, M. Luo, D. S. Brown, D. Seita, B. Thananjeyan, E. Novoseller, and K. Goldberg, "Lazydagger: Reducing context switching in interactive imitation learning," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2021, pp. 502–509.

[25] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer, "Ensembledagger: A bayesian approach to safe imitation learning," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5041–5048.

[26] R. Wang, C. Ciliberto, P. V. Amadori, and Y. Demiris, "Random expert distillation: Imitation learning via expert policy support estimation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6536–6544.

[27] K. Brantley, W. Sun, and M. Henaff, "Disagreement-regularized imitation learning," in *International Conference on Learning Representations*, 2020.

[28] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862*, 2017.

[29] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin, "Primal wasserstein imitation learning," in *International conference on learning representations*, 2021.

[30] K. Brantley, "Expert-in-the-loop for sequential decisions and predictions," Ph.D. dissertation, 2021.

[31] M. Bansal, A. Krizhevsky, and A. S. Ogale, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst," in *Robotics: Science and Systems XV*, 2019.

[32] J. Wong, A. Tung, A. Kurenkov, A. Mandlekar, L. Fei-Fei, S. Savarese, and R. Martín-Martín, "Error-aware imitation learning from teleoperation data for mobile manipulation," in *Conference on Robot Learning*. PMLR, 2021, pp. 1367–1378.

[33] J. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun, "Mitigating covariate shift in imitation learning via offline data with partial coverage," *Advances in Neural Information Processing Systems*, vol. 34, pp. 965–979, 2021.

[34] C.-C. Chuang, D. Yang, C. Wen, and Y. Gao, "Resolving copycat problems in visual imitation learning via residual action prediction," *arXiv preprint arXiv:2207.09705*, 2022.

[35] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] C. Wen, J. Lin, T. Darrell, D. Jayaraman, and Y. Gao, "Fighting copycat agents in behavioral cloning from observation histories," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2564–2575, 2020.

[37] P. Florence and C. Lynch, "Decisiveness in Imitation Learning for Robots," https://ai.googleblog.com/2021/11/decisiveness-in-imitation-learning-for.html, 2022.

[38] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*. PMLR, 2021, pp. 158–168.

[39] Y. Song and D. P. Kingma, "How to train your energy-based models," *arXiv preprint arXiv:2101.03288*, 2021.

[40] S. Russell, "Learning agents for uncertain environments," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 101–103.

[41] B. Piot, M. Geist, and O. Pietquin, "Bridging the gap between imitation learning and inverse reinforcement learning," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 8, pp. 1814–1826, 2016.

[42] B. Lian, W. Xue, F. L. Lewis, and T. Chai, "Robust inverse q-learning for continuous-time linear systems in adversarial environments," *IEEE Transactions on Cybernetics*, 2021.

[43] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[44] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[45] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.

[46] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[47] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[48] F. Torabi, G. Warnell, and P. Stone, "Recent advances in imitation learning from observation," *arXiv preprint arXiv:1905.13566*, 2019.

[49] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 2641–2646.

[50] A. Y. Ng, S. Russell *et al.*, "Algorithms for inverse reinforcement learning." in *Icml*, vol. 1, 2000, p. 2.

[51] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," *Advances in neural information processing systems*, vol. 29, 2016.

[52] K. Kim, S. Garg, K. Shiragur, and S. Ermon, "Reward identification in inverse reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5496–5505.

[53] D. Jarrett, A. Hüyük, and M. Van Der Schaar, "Inverse decision modeling: Learning interpretable representations of behavior," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4755–4771.

[54] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 729–736.

[55] J. Bagnell, J. Chestnutt, D. Bradley, and N. Ratliff, "Boosting structured prediction for imitation learning," *Advances in Neural Information Processing Systems*, vol. 19, 2006.

[56] N. D. Ratliff, D. Silver, and J. A. Bagnell, "Learning to search: Functional gradient techniques for imitation learning," *Autonomous Robots*, vol. 27, no. 1, pp. 25–53, 2009.

[57] U. Syed and R. E. Schapire, "A game-theoretic approach to apprenticeship learning," *Advances in neural information processing systems*, vol. 20, 2007.

[58] N. Aghasadeghi and T. Bretl, "Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1561–1566.

[59] M. Kalakrishnan, P. Pastor, L. Righetti, and S. Schaal, "Learning objective functions for manipulation," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1331–1336.

[60] A. Boularias, J. Kober, and J. Peters, "Relative entropy inverse reinforcement learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 182–189.

[61] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *arXiv preprint arXiv:1507.04888*, 2015.

[62] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *International conference on machine learning*. PMLR, 2016, pp. 49–58.

[63] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning." in *IJCAI*, vol. 7, 2007, pp. 2586–2591.

[64] J. Choi and K.-E. Kim, "Map inference for bayesian inverse reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[65] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear inverse reinforcement learning with gaussian processes," *Advances in neural information processing systems*, vol. 24, 2011.

[66] D. Brown, R. Coleman, R. Srinivasan, and S. Niekum, "Safe imitation learning via fast bayesian reward inference from preferences," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1165–1177.

[67] A. J. Chan and M. van der Schaar, "Scalable bayesian inverse reinforcement learning," in *International Conference on Learning Representations*, 2021.

[68] D. Lindner, A. Krause, and G. Ramponi, "Active exploration for inverse reinforcement learning," *arXiv preprint arXiv:2207.08645*, 2022.

[69] A. M. Metelli, G. Ramponi, A. Concetti, and M. Restelli, "Provably efficient learning of transferable rewards," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7665–7676.

[70] J. Ho, J. Gupta, and S. Ermon, "Model-free imitation learning with policy optimization," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2760–2769.

[71] S. Levine and V. Koltun, "Continuous inverse optimal control with locally optimal examples," *arXiv preprint arXiv:1206.4617*, 2012.

[72] A. Deka, C. Liu, and K. P. Sycara, "Arc-actor residual critic for adversarial imitation learning," in *Conference on Robot Learning*. PMLR, 2023, pp. 1446–1456.

[73] R. Bhattacharyya, B. Wulfe, D. J. Phillips, A. Kuefler, J. Morton, R. Senanayake, and M. J. Kochenderfer, "Modeling human driving behavior through generative adversarial imitation learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 2874–2887, 2023.

[74] J. Song, H. Ren, D. Sadigh, and S. Ermon, "Multi-agent generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 31, 2018.

[75] M. Orsini, A. Raichuk, L. Hussenot, D. Vincent, R. Dadashi, S. Girgin, M. Geist, O. Bachem, O. Pietquin, and M. Andrychowicz, "What matters for adversarial imitation learning?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 656–14 668, 2021.

[76] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," in *International Conference on Learning Representations*, 2018.

[77] I. Kostrikov, K. K. Agrawal, D. Dwibedi, S. Levine, and J. Tompson, "Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning," in *International conference on learning representations*, 2019.

[78] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[79] Y. Li, J. Song, and S. Ermon, "Infogail: Interpretable imitation learning from visual demonstrations," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[80] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[81] A. Edwards, H. Sahni, Y. Schroecker, and C. Isbell, "Imitating latent policies from observation," in *International conference on machine learning*. PMLR, 2019, pp. 1755–1763.

[82] Y. Hu, G. Chen, Z. Li, and A. Knoll, "Robot policy improvement with natural evolution strategies for stable nonlinear dynamical system," *IEEE Transactions on Cybernetics*, 2022.

[83] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1118–1125.

[84] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," in *Imitation, Intent, and Interaction (I3) Workshop at ICML 2019*, June 2019.

[85] A. J. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, vol. 2. IEEE, 2002, pp. 1398–1403.

[86] S. Choe, H. Seong, and E. Kim, "Indoor place category recognition for a cleaning robot by fusing a probabilistic approach and deep learning," *IEEE Transactions on Cybernetics*, 2021.

[87] D. S. Raychaudhuri, S. Paul, J. Vanbaar, and A. K. Roy-Chowdhury, "Cross-domain imitation from observations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8902–8912.

[88] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[89] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, p. 4950–4957.

[90] T. Gangwani, Y. Zhou, and J. Peng, "Imitation learning from observations under transition model disparity," in *International Conference on Learning Representations*, 2022.

[91] A. Jaegle, Y. Sulsky, A. Ahuja, J. Bruce, R. Fergus, and G. Wayne, "Imitation by predicting observations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4665–4676.

[92] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[93] W. Sun, A. Venkatraman, G. J. Gordon, B. Boots, and J. A. Bagnell, "Deeply aggrevated: Differentiable imitation learning for sequential prediction," in *International conference on machine learning*. PMLR, 2017, pp. 3309–3318.

[94] M. Pflueger, A. Agha, and G. S. Sukhatme, "Rover-irl: Inverse reinforcement learning with soft value iteration networks for planetary rover path planning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1387–1394, 2019.

[95] W. Jeon, S. Seo, and K.-E. Kim, "A bayesian approach to generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[96] F. Sasaki, T. Yohira, and A. Kawaguchi, "Sample efficient imitation learning for continuous control," in *International conference on learning representations*, 2018.

[97] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," in *International conference on machine learning*. PMLR, 2019, pp. 783–792.

[98] J. Zhang, R. Zhu, and E. Ohn-Bar, "Selfd: Self-learning large-scale driving policies from the web," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 316–17 326.

[99] Y.-H. Wu, N. Charoenphakdee, H. Bao, V. Tangkaratt, and M. Sugiyama, "Imitation learning from imperfect demonstration," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6818–6827.

[100] F. Sasaki and R. Yamashina, "Behavioral cloning from noisy demonstrations," in *International Conference on Learning Representations*, 2020.

[101] Y. Wang, C. Xu, B. Du, and H. Lee, "Learning to weight imperfect demonstrations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 961–10 970.

[102] G.-H. Kim, S. Seo, J. Lee, W. Jeon, H. Hwang, H. Yang, and K.-E. Kim, "Demodice: Offline imitation learning with supplementary imperfect demonstrations," in *International Conference on Learning Representations*, 2022.

[103] M. Beliaev, A. Shih, S. Ermon, D. Sadigh, and R. Pedarsani, "Imitation learning by estimating expertise of demonstrators," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022, pp. 1732–1748.

[104] J. Chae, S. Han, W. Jung, M. Cho, S. Choi, and Y. Sung, "Robust imitation learning against variations in environment dynamics," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162. PMLR, 2022, pp. 2828–2852.

[105] B. C. Stadie, P. Abbeel, and I. Sutskever, "Third-person imitation learning," in *International conference on learning representations*, 2017.

[106] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi, "Xirl: Cross-embodiment inverse reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 537–546.

[107] A. Fickinger, S. Cohen, S. Russell, and B. Amos, "Cross-domain imitation learning via optimal transport," in *International Conference on Learning Representations*, 2022.

[108] K. Kim, Y. Gu, J. Song, S. Zhao, and S. Ermon, "Domain adaptive imitation learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5286–5295.

[109] M. Yang, S. Levine, and O. Nachum, "Trail: Near-optimal imitation learning with suboptimal data," in *International Conference on Learning Representations*, 2022.

[110] G. Xiang and J. Su, "Task-oriented deep reinforcement learning for robotic skill acquisition and control," *IEEE transactions on cybernetics*, vol. 51, no. 2, pp. 1056–1069, 2019.

[111] R. Ramrakhya, E. Undersander, D. Batra, and A. Das, "Habitat-web: Learning embodied object-search strategies from human demonstrations at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5173–5183.

[112] J. Lee, W. Jeon, B. Lee, J. Pineau, and K.-E. Kim, "Optidice: Offline policy optimization via stationary distribution correction estimation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6120–6130.

[113] W. Sheng, A. Thobbi, and Y. Gu, "An integrated framework for human–robot collaborative manipulation," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2030–2041, 2015.