CG-001

バイタルデータと投与情報を用いた術中管理支援のための逆強化学習に基づくイベント予測

Event prediction based on inverse reinforcement learning for intraoperative care support using vital data and administration information

角 文真 † 濱上 知樹 †
眞一 弘士 § 増井 健一 § 大嶽 浩司 §
Fumimasa SUMI Tomoki HAMAGAMI
Hiroshi MAKAZU Kenichi MASUI Hiroshi OTAKE

1. はじめに

IoT 等で収集される多次元時系列データが増加し，その解析・活用の試みも拡大している．一方，医療分野では麻酔科医不足が深刻化しており，常勤麻酔科医への過度な負担集中，必要となる手術施行への影響といった問題が生じている．麻酔科医の主な業務は手術患者の術中管理であり，これは知識や経験に基づき行われている．そこで，本研究では術中管理を対象とし，多次元時系列データの解析・活用を試みる．具体的には，術中管理のログデータである麻酔記録を活用し，必要な措置を予測・通知する支援システムの実現を目指す．

術中管理は高次元かつ非線形な性質を持つタスクであるため，術中管理則の明示的な付与は困難である．そこでデータドリブンな手法である模倣学習に着目する．模倣学習には教師あり学習と逆強化学習が存在するが，教師あり学習と比較し，必要データが少なく長期予測にも有効とされる逆強化学習を用いる．特に，GAN (Generative Adversarial Networks) の枠組みで逆強化学習を行う手法が動画予測等で高い成果を上げており，その一種である AIRL (Adversarial Inverse Reinforcement Learning)[1] を用いる．しかし，術中管理に適用するには 2 つの問題がある．まず，術中管理は測定バイタル，使用投与物が多く，状態行動空間が広大となる．次に，周術期リスクの低い患者では，麻酔科医の措置を要する場面が少なく，術野とバイタルの監視が多くの時間を占める．ゆえに，AIRL の報酬関数が「何もしていない状態」を麻酔科医らしさとして学習し，措置の予測が困難になることが考えられる．

そこで 1 点目については VAE (Variational Autoencoder)[2] による次元圧縮を導入する．2 点目については必要措置実行の予測成功時に報酬を付与する報酬設計をする．この 2 つを併用する AIRL を提案システムとし，麻酔記録を用い f 値で評価する．

2. 術中管理と麻酔記録

全身麻酔や局所麻酔下の手術患者の循環・呼吸・意識 (疼痛) を監視し，薬剤使用等で安全維持することを術中管理という．

麻酔科医の業務の中で最も一般的なものが，全身麻酔下で行う手術の術中管理である．全身麻酔は鎮静 (意識消失)，鎮痛，筋弛緩，有害反射抑制の 4 要素に作用する薬剤を用いる．かつては吸入麻酔薬単剤で 4 要素をある程度満たす方法で行われたが，現在は各要素に特異的に作用する薬剤を用い，それぞれに対し個別に十分な制御を行うバランス麻酔が主流である．全身麻酔のための薬剤投与の他、体液管理や呼吸管理等も患者の循環・呼吸・意識の安定化のため行われる．

このような麻酔科医により術中に行われることのほぼすべてと患者のバイタルデータが時系列で記録されたものを麻酔記録という．

3. Adversarial Inverse Reinforcement Learning

本研究で使用する AIRL は GAN の枠組みで逆強化学習を行う手法の 1 つである．AIRL では，現在の方策 $\pi(a|s)$ のもとで軌跡 $\tau = (s_0, a_0, ..., s_T, a_T)$ を生成し，エキスパート軌跡とともに discriminator へ入力する．discriminator は状態行動対がエキスパートのものかエージェントのものか識別するように訓練し，discriminator の出力を基に方策を更新する．類似モデルに GAIL (Generative Adversarial Imitation Learning)[3] が存在するが，直接報酬関数を推定せず，方策の推定のみを行う点が異なる．

逆強化学習には同じ最適方策が生じる報酬関数が複数存在する報酬設計問題がある．これを解決するため，AIRL では識別関数にパラメータ θ_{airl} と ϕ_{airl} を導入し，以下のような識別関数を設計する．最適化を行うと $f_{\theta_{airl},\phi_{airl}}$ に報酬関数が学習される．これにより，報酬関数の取り出しも可能となる．

$$D_{\theta_{airl},\phi_{airl}}(s,a,s') = \frac{\exp\{f_{\theta_{airl},\phi_{airl}}(s,a,s')\}}{\exp\{f_{\theta_{airl},\phi_{airl}}(s,a,s')\} + \pi_\theta(a|s)} \tag{1}$$

AIRL の方策の更新は方策勾配法の一種である Trust Region Policy Optimization (TRPO)[4] で行う．TRPO は以下の制約付き最適化問題を解くことで確率的方策 π_θ のパラメータを更新する．

$$\underset{\theta}{maximize}\, L_{\theta_{old}}(\theta) = \mathbb{E}_{s\sim\rho_{\theta_{old}},a\sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)} Q_{\theta_{old}}(s,a)\right] \tag{2}$$

$$subject\,to\, \bar{D}_{KL}(\theta_{old},\theta) = \mathbb{E}_{s\sim\rho_{\theta_{old}}}[D_{KL}(\pi_{\theta_{old}}(\cdot|s)||\pi_\theta(\cdot|s))] \leq \delta \tag{3}$$

ここで ρ は初期状態分布，a は行動，s は状態，s' は次の状態，δ はステップサイズ，θ_{old} は方策更新前のパラメータ，θ は方策更新後のパラメータ，$Q_{\theta_{old}}$ は方策更新前の行動価値関数である．強化学習では大幅な更新を行い方策が悪化した場合，悪化した方策の下で次のバッチが収集される．TRPO は KL ダイバージェンスを一定値 δ 以下に抑える更新幅での期待割引報酬最大化で，この問題を解決している．

4. Variational Autoencoder

本研究で次元圧縮手法として用いた VAE は深層生成モデルの 1 つである．次式の最大化で，入力データ $\mathbf{x}$ を潜在変数 $\mathbf{z}$ に次元圧縮して特徴量を獲得すると共に，潜在変数を正規分布

† 横浜国立大学大学院理工学府
§ 昭和大学病院

化する学習を行う．

$$\mathcal{L}(\theta_{vae}, \phi_{vae}, \mathbf{x}) = -D_{KL}(q_{\phi_{vae}}(\mathbf{z}|\mathbf{x})||p_{\theta_{vae}}(\mathbf{z})) + \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] \quad (4)$$

ここで，式の第一項は正則化項：KL ダイバージェンスであり，第二項は入力データと出力データの復元誤差の項である．

5. 提案システム

5.1 イベント予測と目標

本研究ではイベントを麻酔科医による投与物の投与量の変更と定義する．点滴や持続投与などの一定速度で投与を続ける投与は投与量変更時刻を，単回投与などのその時刻のみ投与が実施するものは投与実施時刻をイベントとして扱う．イベント予測では，イベントの発生時刻，イベントの種類，イベントの程度といった予測が考えられる．後者二つは本研究では投与物の種類，変更される投与量である．

最終的な目標は回帰によるこれらの予測であるが，本研究では現時刻より一定時間内のイベント発生投与物を予測する．

5.2 状態と行動の定義

麻酔科医は，患者のコンディションを良好に維持するために術中管理を行う．しかし，患者のコンディションを状態，投与を行動とすることは困難である．これは，バイタルと投与の関係は高次元かつ非線形であり，モデル作成により投与に対するバイタルの変化をシミュレートすることが困難なためである．

そこで，本研究では，行動が確定した場合に状態遷移が決定論的になる状態と行動を定義し，術中管理の遷移則を方策で獲得する．状態を時刻 $t-\tau_s \sim t$ の観測，行動を時刻 $t+1 \sim t+1+\tau_a$ の観測と設定することでこれを可能とした．ここで τ は任意の数値であり，状態と行動それぞれの観測に対して設定した時間窓から 1 を引いた値となる．観測は 6.1 節で述べるようにバイタル，投与，体重である．

5.3 AIRL を用いたシステム

5.3.1 関連研究

逆強化学習による予測の研究に GAIL を用いた動画の予測 [5] がある．未来のフレーム生成，60 フレーム後における行動の識別等の実験で LSTM 等の従来手法を上回る精度を出し，GAN の枠組みで行う逆強化学習の有効性が示されている．

5.3.2 システム概要

本研究では図 1 に示すシステムでエキスパート方策を回復し，イベント予測を行う．まず，事前訓練された VAE の encoder により次元圧縮された観測データで構成されたエキスパート軌跡を AIRL 与える．エキスパート軌跡よりランダムに選択された初期状態より，現在の方策 π に従いエージェント軌跡を生成する．エキスパート状態行動対とエージェント状態行動対を discriminator に入力し，報酬を得る．同時に，初期状態よりエキスパート方策に従う場合の未来の観測と現在の方策に従う場合の未来の観測に含まれるイベントを比較し，その結果に応じて報酬を与える．得られた報酬を基に方策を更新する．運用時は，図 2 のように観測されたデータを次元圧縮し，方策に基づき未来の観測を予測する．予測された未来の観測に含まれるイベントを識別し，発生が予測されるイベントを麻酔科医に通知する．本研究では，AIRL 単体ではなくイベント予測を行うに際し，特徴空間の探索とイベント報酬設計の 2 つを併用し，これを提案システムとして評価を行った．

5.3.3 特徴空間の探索

観測は 6.1 節で述べるように全 117 次元である．しかし，麻酔科医が常時このすべてを認識し，術中管理を遂行するとは考えにくい．例えば，6.1 節で述べるように投与情報は全 48 次元であるが，ある時刻に同時に使用されるものは多くとも 11 である．この際，麻酔科医は使用中の投与物のみを認識すると考えられる．そこで，Zheng らの実験 [5] 同様に状態空間・行動空間の縮小を目的に次元圧縮を導入した．

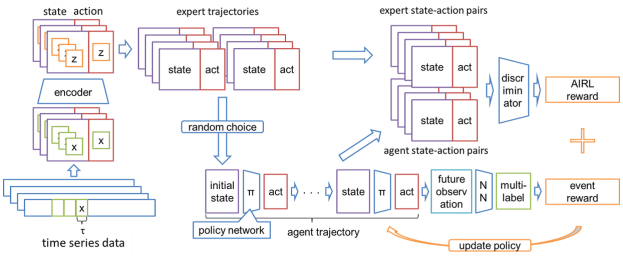

図 1 システム全体構成 (学習時)

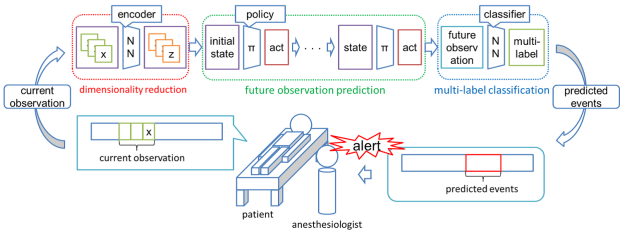

図 2 システム全体構成 (運用時)

提案システムでは，次元圧縮で獲得した観測の特徴量を AIRL に入力し，学習する．獲得した方策で軌跡を生成して未来の観測を予測する．予測した観測に事前訓練された multi-label 分類器を用い，含まれるイベントを識別する．

代表的な次元圧縮手法には VAE の他に Autoencoder(AE) があるが，VAE は AE と異なり特徴量が正規分布となるため，探索がより効率的になると考えられ採用した．

ニューラルネットワーク (NN) による multi-label 分類は，潜在変数から直接イベントを識別する方法として採用した．提案システムでは未来の観測の特徴量を予測する．そのため，decoder で観測の復元を行いイベントを識別する方法は非効率であるためこの方法を採用した．

5.3.4 イベント予測報酬設計

本研究では投与量の変更をもってイベントとする．術中管理では頻繁な投与量の変更は起きず，特定の投与物でイベントが発生するケースは極めて稀となる．ゆえに投与量の変化をエキスパートらしさとみなさない報酬関数が獲得される可能性が高い．この場合，投与量が変動する状態行動対に低い報酬を与えるため，エキスパート軌跡の復元が困難となる．そこで，本研究ではイベント予測に成功した場合に報酬を与えることでこの問題の解決を試みた．バイタルの変動等を含め，実際に起こりうる観測が生成されるよう制約する非線形で設計困難な報酬の設計を AIRL の報酬関数が担い，一方でイベント予測報酬でイベントが発生する観測が生成されるようにする．この 2 種類の報酬関数でエキスパート方策の回復を行う．

イベント予測報酬の与え方を図 3 に示す．初期状態からエージェント方策に従うことでサンプリングされる未来の観測に対し，NN で含まれるイベントを識別する．エキスパート方策に従った場合の未来の観測にイベントが含まれる場合は，エージェント方策に従った場合の未来の観測に含まれるイベントと比較し，その結果に応じて報酬を与える．イベントが含まれない場合は報酬は与えない．

6. 実験

6.1 使用データ

本研究ではフリーの電子麻酔記録ソフト paperChart で記録された 499 件の実際の手術の麻酔記録を使用する．バイタ

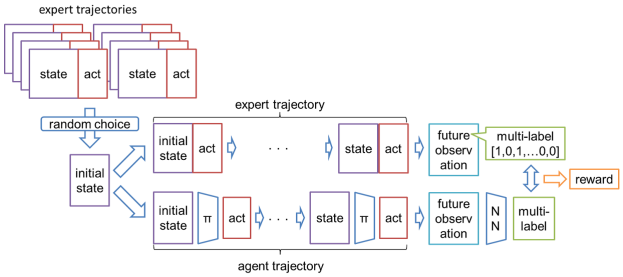

図 3　イベント予測に基づく報酬設計

ルデータの指標は表 1 に示す 17 項目である．各項目について，一分間における平均値，中央値，最大値，最小値が測定されており，バイタルデータは 68 次元で構成される．投与情報は，投与物名とその投与方法で区別を行った．スケールの大きく異なる値を含む場合は別の区分として扱い，表 2 に示す 48 区分となった．ここで，投与物名は paperChart における登録名に準じている．投与物名中の?は区切り文字であり，その前の数字は日本麻酔科学会規定の薬剤コードである．投与方法は B は単回投与，P は持続投与，D は点滴投与，G はガーゼ，S は吸引を示す．表 2 に 2 つあるアセリオ/B の片方は半角カタカナで登録されており，2 つは区別されている．時間粒度はバイタルデータの時間粒度 1 分に統一した．これに患者の体重を加えた計 117 次元で 1 分における観測データが表される．

前処理として，データの欠損値に関しては最後に観測された値で補完し，各次元ごとに正規化を行った．提案手法の訓練には 400 件，テストには 99 件を割り当てている．

表 1　バイタルデータ測定項目

測定項目	
心拍数	吸気酸素濃度
心室性期外収縮（不整脈）	呼気酸素濃度
非観血的血圧 (s)	吸気亜酸化窒素濃度
非観血的血圧 (m)	呼気亜酸化窒素濃度
非観血的血圧 (d)	吸気セボフレン濃度
抹消血中酸素飽和度	呼気セボフレン濃度
心電図 ST セグメント	呼気数
吸気二酸化炭素	体温
呼気二酸化炭素	

6.2　実験設定

6.2.1　共通設定と各設定

表 3 に示す 3 種類の設定の下実験を行った．共通の設定として，時刻 t の状態 s_t を時刻 t 以前の 12 分間の観測 $obs_{t-11\sim t}$ とし，12 分後までの観測 $obs_{t+1\sim t+12}$ を予測した．予測結果に multi-label 分類器を適用してイベントの識別を行い，f 値で評価した．イベント予測報酬は f 値× 1000 とした．

AIRL の各設定ではイベント発生タイミングの不確実性を考慮している．麻酔科医の行動は分単位の場合が多いため，イベント発生時刻の前後 1，2 分ずれが処置としては問題ない可能性がある．すなわち，イベントの発生自体は予測されてもタイミングを特定する特徴がなく，予測の難易度が上がる可能性がある．各設定ではこの点を考慮し，直後の観測のみを予測する AIRL3，より不確実性を考慮し一度の予測範囲を 4 分とした AIRL2，時間軸方向の次元圧縮で，時間軸方向の変化の特徴を扱う AIRL1 を設定した．AIRL1 では投与量変化の特徴を扱うため，イベント発生タイミングはより厳密でなくなる．

6.3　比較と評価

対照実験を 2 種類行い提案手法を評価した．既存手法である教師あり学習との比較，AIRL1 に対し異なる報酬設計をした場合の比較を行った．

表 2　使用投与物とその投与方法

投与物/投与方法	
O₂/P	ガーゼ/G
0105?N₂O/P	CEZ/B
0101?Sev/P	CEZ/D
P140/D	0501?ネオシネジン/B
air/P	0401?0.375% アナペイン (硬)/B
0201?フェンタニル/B	セフメタゾール/B
0516?硫アト/B	セフメタゾール/D
アセリオ/B	アセリオ/B
吸引量/S	デキサート/B
尿量/S	0405?脊マーカイン (等)/B
BC/D	アセリオ 100ml/B
0106?プロポフォール/B	アセリオ 100ml/D
0106?プロポフォール/P	ペンタジン/B
0306?Rb/B	ネオシネジン持続/P
抗菌薬/B	0518?ニカルピン/B
NS/D	0.1875% アナペイン (局)/B
0202?アルチバ/B	RCC/D
0202?アルチバ/P	1%E キシロカイン (局)/B
ブリディオン/B	0401?0.375% アナペイン (局)/B
0102?Des/P	T₁/D
ラクテック/D	0401?0.25% アナペイン (硬)/B
0501?エフェドリン/B	0406?脊マーカイン (高)/B
自己血/D	吸引量 2/S
ロピオン/B	尿量 2/S

表 3　AIRL の設定

設定	行動 (分)	圧縮単位 (分)	episode 長
AIRL1	4	4	3
AIRL2	4	1	3
AIRL3	1	1	12

6.3.1　既存手法との比較

比較対象として教師あり学習を用いた回帰，回帰は行わずイベント予測のみを行う分類を採用した．50 回学習を行い精度の中央値と最大値を調査した．モデルとしては NN を使用する．NN による回帰の各設定は表 4 に示す．f 値については，未来の観測にイベントが含まれないエピソードも含め，予測された未来の観測に multi-label 分類器を適用し，イベント識別結果の f 値を評価指標として用いる．

表 4　NN 回帰の設定

設定	行動 (分)	圧縮単位 (分)	episode 長
NN 回帰 1	4	4	3
NN 回帰 2	4	1	3
NN 回帰 3	1	1	12
NN 回帰 VAE なし 1	4	なし	3
NN 回帰 VAE なし 2	1	なし	12

6.3.2　報酬設計の比較

AIRL1 に対し報酬設計の対照実験を実施した．AIRL 報酬単独時とイベント報酬単独時を比較対象とした．前者はイベント予測報酬の効果，後者は AIRL 報酬の効果の検証を目的とする．イベント報酬単独時は，コスト (負の報酬) として (1-f 値) × (-1000) を与える，この時 f 値は，イベントが発生する系列はイベント発生予測，イベントが発生しない系列は未発生予測に対するものである．評価は f 値と未来の観測に対する回帰の二乗平均平方根誤差 (rmse) とした．

7. 結果と考察

7.1 既存手法との比較

7.1.1 各設定の f 値

各設定における f 値は表 5 のようになる．学習進行状況より AIRL1 と AIRL3 は 20000 イテレーション，AIRL2 は 17500 イテレーション時の f 値を採用した．NN は中央値と括弧内に最高精度を記載している．ベースラインとして NN 回帰を，参考としてイベントの予測のみを行った NN 分類を示している．

AIRL1 がテストデータに対し f 値 0.307 とベースライン

表 5 各設定の f 値

項目	訓練データ (f 値)	テストデータ (f 値)
AIRL1	0.358	0.307
AIRL2	0.377	0.283
AIRL3	0.418	0.257
NN 回帰 1	0.239 (0.270)	0.219 (0.248)
NN 回帰 2	0.186 (0.213)	0.196 (0.220)
NN 回帰 3	0.149 (0.178)	0.159 (0.187)
NN 回帰 VAE なし 1	0.100 (0.233)	0.081 (0.225)
NN 回帰 VAE なし 2	0.0 (0.177)	0.0 (0.173)
NN 分類 1	0.287 (0.450)	0.201 (0.308)
NN 分類 2	0.563 (0.724)	0.267 (0.292)
NN 分類 VAE なし	0.334 (0.425)	0.268 (0.311)

と比較して最高精度であり，ベースラインの最高精度から f 値 0.059 の改善となった．逆強化学習では報酬の伝搬により未来の状態まで考慮した方策が獲得される．教師あり学習は行動のみを考慮した方策を獲得するため，提案手法がベースラインを超えたと考えられる．また，提案手法の精度は参考値の学習 50 回中の中央値を上回り最高精度と同程度となった．これについて 2 点理由が考えられる．

まず，TRPO では方策の更新幅を制限し，方策が改善する範囲で更新を行う．教師あり学習では方策が悪化した場合，次のバッチの学習にて修正を行うため，TRPO を用いた提案手法の方が学習が安定すると考えられる．次に，提案手法では報酬により f 値を直接最大化する．教師あり学習では binary cross entropy の最小化によって multi-label に対する尤度最大化で f 値を最大化するため，提案手法の方が f 値向上の点で優位性があると考えられる．

テストデータにおいて AIRL1〜3 では AIRL1 が，NN 回帰 1〜3 では NN 回帰 1 が最高精度となった．いずれも時間軸方向の次元圧縮を加えた設定であり，6.2.1 項にて述べたイベント発生タイミングの不確実性の考慮の有効性が示された．

次元圧縮は NN 回帰では f 値を向上させている．一方，分類では NN 分類 1 の f 値の中央値は VAE なしを 0.067 下回るが，最高精度や AIRL1 では f 値が同程度であるため，次元圧縮による情報欠落の精度への影響は少ないと考えられる．

7.1.2 投与物別の f 値

精度の高かった AIRL1 の投与物別の f 値を図 4 に示す．全体の傾向としてイベント数と f 値が比例関係にあるが，イベント数に反して精度が高いもの，低いものが存在する．これは状態 (観測 12 分) における特徴の有無が一因と考えられる．

7.2 報酬設計の比較

20000 イテレーション時の AIRL1 における報酬設計の比較の結果を表 6 に示す．イベント報酬の有無はテストデータに対する f 値で評価した．AIRL 報酬単独時は f 値が 0.072 と低いことより，AIRL 報酬関数はイベントをエキスパートらしさと捉えず，イベント予測報酬を要することが示された．AIRL 報酬の有無は未来の観測に対する回帰を rmse で評価した．イベント報酬単独時は rmse が 2.904 と大きい．強化学習は f 値

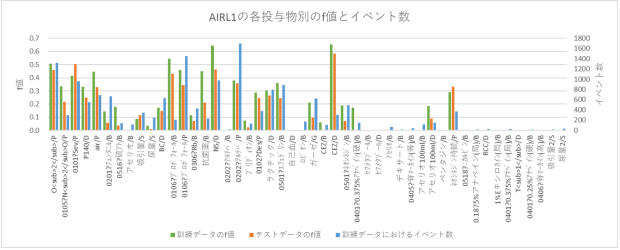

図 4 投与物別の f 値と訓練データにおけるイベント数

最大化のみを行うことから，現実には起こりえない状態遷移も発生するためと考えられる．これら 2 つの結果より，提案システムの報酬設計の意図通り，AIRL 報酬で実際に起こりうる状態遷移に制約しつつイベント予測を行えていると考えられる．

表 6 各報酬の有無と f 値

設定	f 値	rmse
提案手法	0.307	0.400
AIRL 報酬単独	0.072	0.322
イベント報酬単独	0.291	2.904

8. おわりに

本研究では AIRL に VAE による次元圧縮とイベント予測報酬を加えたシステムを提案し，現時刻から一定時間内のイベント発生投与物を予測した．これを既存手法と比較したところ，提案システムが高精度となった．このことから，提案システムによるイベント予測が可能と示された．

提案システムによるイベント予測精度はベースラインを上回り，イベント分類の最高精度と同程度であったが，f 値は 0.3 程度であった．イベント数が多いにもかかわらず，f 値の低い投与物も存在することから，特徴抽出に課題があると考えられる．今後は，状態を観測 12 分以上の長期系列とすることや，イベント予測報酬設計の改善により精度の向上を目指す．

参考文献

[1] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *Proceedings of the 6th International Conference on Learning Representations*. 2018.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*. 2014.

[3] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, pp. 4565–4573. 2016.

[4] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Vol. 37*, pp. 1889–1897. ICML, 2015.

[5] Kuo-Hao Zeng, William B. Shen, De-An Huang, Min Sun, and Juan Carlos Niebles. Visual forecasting by imitating dynamics in natural sequences. In *IEEE International Conference on Computer Vision, ICCV 2017*, pp. 3018–3027, 2017.