# 任意の目的地に対する Map-Free 強化学習ナビゲーションの安定化に関する研究

○天野 大輔（明治大学）　森岡 一幸（明治大学）

This paper proposes a map-free navigation method for mobile robots using deep reinforcement learning with visual and geometric observations. A novel start and goal configurations enable training including several different situations. Furthermore, an angular feature indicating the global direction from the start to the goal is introduced to improve policy stability as one of the observations. The effect of incorporating LSTM is also evaluated. The performances of the proposed method are experimentally evaluated in simulation environments.

## 1. 緒言

近年，深層強化学習を用いた行動モデルによる移動ロボットの自律ナビゲーションのの研究がいくつか行われている．筆者らは，行動モデルへの外界センサーからの入力情報として，走行環境をセグメンテーションした単眼カメラ画像を用いた手法 [1] や，深度カメラによる距離データを用いた手法 [2] を提案してきた．深度カメラは Meta で開発されたシステム [3] でも行動モデルによるナビゲーションに使用されている．これらの入力情報は，学習環境と実環境のギャップが小さく，Sim2Real を実現する上で有効である．

本研究は，セグメンテーションした単眼カメラ画像に基づく，Map-Free なナビゲーションシステムを実現することを目指している．上述した研究例ではいずれも何らかの手法で取得した自己位置と目的地との相対距離・姿勢情報を合わせて入力する行動モデルとなっている．視覚情報に強く依存する構成であり，奥行き推定や方向認識が不安定であり，目的地によっては到達することが困難である．本稿では，このような背景を踏まえ，視覚情報に加えて補助的な観測量を導入することで，目的地の柔軟な設定に対応可能な Map-Free ナビゲーションのための行動モデルの学習に関して報告する．

## 2. 強化学習環境の構成

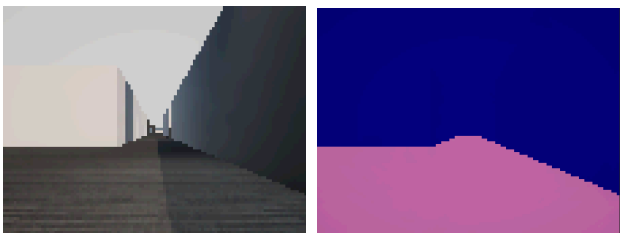
本研究では，著者らが所属する大学の廊下を模したシミュレーション環境を Unity 上に構築し，移動ロボットに対する視覚強化学習モデルの訓練を行った．本環境は，複数の交差点や分岐，長い廊下や隘路などを含む複雑な構造を有しており，遠距離目標への回り込みやすれ違いを伴うような高度なナビゲーション行動も求められる．

本章では，任意の目的地への汎用的なナビゲーション方策の獲得を目的とし，強化学習環境の構成要素として，視覚観測と動作出力の設計，訓練用マップの構造，および初期位置と目的地の生成方式について詳述する．

### 2.1 エージェントの設計

#### 2.1.1 エージェント視点のカメラ設定

本研究では，実機ロボットにて単眼カメラを搭載することを想定したエージェントを用いる．カメラはエージェントの前面に固定し，地面からの高さは約 77 cm，水平方向に対して下方に約 4 度傾けている．視野角は垂直方向に 60 度とし，前方の床面や障害物の広い範囲を撮影できる．取得画像の例を図 1（A）に示す．



（A）オリジナル画像　　（B）二値化後の視点画像

図 1　エージェントの視点画像例

#### 2.1.2 エージェントの動作設計と制御方式

エージェントの移動動作として並進速度と角速度の連続値を方策ネットワークから出力し，それぞれ上限となる最大速度を定める．加速度には物理的な滑らかさを再現するための制限を設けて，実環境における挙動に近づくよう，速度変化を段階的に補間する．さらに，動作の多様性と過学習を抑制するため，角速度にランダムノイズを導入する．これは，実機ロボットの走行中に生じる微小な振動を模擬することを目的としており，非停止時に微小なゆらぎを加えることで，実環境での頑健性の向上が期待される．位置と姿勢は，更新後の並進速度および角速度に基づいて，差動ロボットの理想的な運動モデルに従って逐次更新される．

### 2.2 シミュレーション環境の設計

#### 2.2.1 学習環境の二値化処理

視覚入力に含まれる不要な色情報や照明条件の変動は，強化学習での不安定化を引き起こす．この影響を抑えるため，床面を赤紫色（RGB: 225,105,180），壁面を暗青色（RGB: 0,0,128）で塗り分けた二値化環境を導入し，実環境での走行時に床面のセマンティックセグメンテーションで二値化することを想定した構成とする [1]．この処理により，視覚センサーからの入力画像におけるばらつきが抑えられ，学習の安定性と方策獲得の効率が向上することが期待される．図 1（B）はエージェント視点画像の一例を示す．

#### 2.2.2 エージェントの初期位置と目的地の生成方法

学習時のエージェントの初期位置および目的地は，任意の目的地に対するナビゲーション方策の汎化性能と訓練時の効率性を高めるため，図 2 に示すような処理により設定する．

図 3 に示された訓練用マップは，6 つのサブマップ

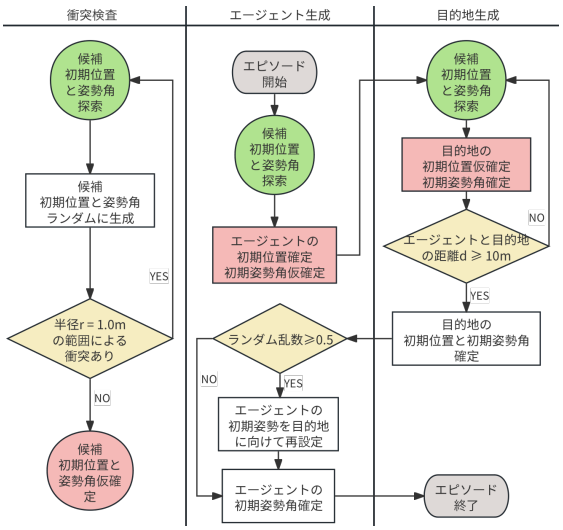



図 2　エージェントの初期位置と目的地の生成方法

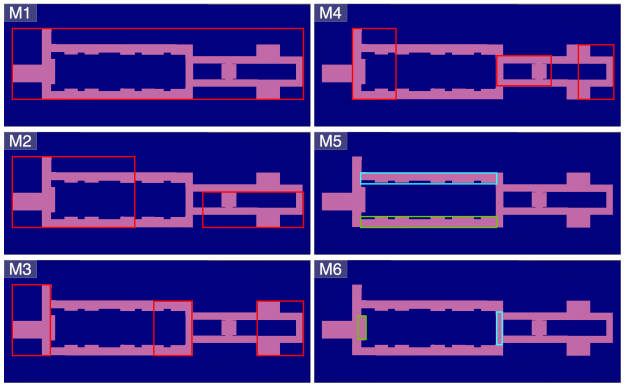



図 3　訓練用マップの分割例

(M1〜M6) に分割して学習に用いる．同時に 13 台のエージェントを効率的に訓練することが可能である．M1〜M4 では，エージェントと目的地の双方を赤色の矩形領域内からランダムに生成する．M1 はマップ全体を対象しており，M2〜M4 では訓練の安定性を高めるため，異なる地形に対応した複数のサブ領域に分割し，特定の構造に対する方策の獲得を促す設計とする．M5 および M6 では，エージェントは緑色の矩形領域から，目的地は水色の矩形領域から独立にランダム生成され，いずれも中央に長い廊下を含むような複雑な地形における遠回り経路の学習を目的としている．いずれも，タスクが過度に単純化されることを防ぐため，エージェントと目的地のユークリッド距離が 10 m 以上となるようにする．また，他物体が存在しない位置に設定するため，半径 $r = 1.0\,\mathrm{m}$ の衝突判定を行い，衝突のない位置が見つかるまで繰り返す．さらに，初期化後，50% の確率でエージェントの姿勢を目的地方向に向けて再設定する．これは，学習初期の探索の難易度を軽減し，目標への移動を促すことを目的としている．

## 3.　強化学習モデルの設計

本章では，視覚入力と IMU を模した観測情報に基づき，エージェントが任意の目的地に自律移動するための強化学習モデルの設計について述べる．到達距離や動作の安定性などを意識した報酬設計を行い，ナビゲーション行動の最適化を目指している．

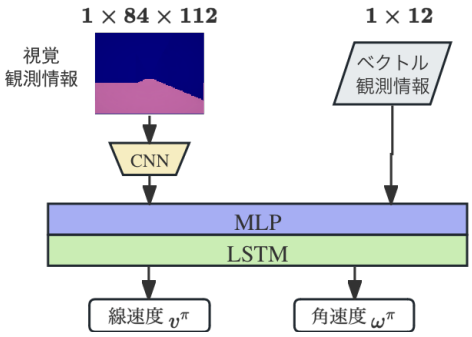



図 4　LSTM 構造付き方策ネットワーク

### 3.1　モデルの構造

強化学習アルゴリズムとして PPO（Proximal Policy Optimization）[4] を採用し，観測情報に基づき連続行動 $a = (v_\pi, \omega_\pi)$ を出力する方策ネットワークを構築する．

また，部分可観測性への対応として，ネットワーク内部に LSTM（Long Short-Term Memory）を導入した．視覚画像とベクトル観測を統合し，LSTM により時系列情報を保持することで，遮蔽環境でも頑健な方策が得られる．このような構成は，部分可観測環境における強化学習に有効であることが報告されている [5]．方策ネットワーク全体の構成を図 4 に示す．

### 3.2　観測情報の設計

#### 3.2.1　全体方位角の導入

著者らの過去の研究では，相対距離や方位角など短期的な観測に基づく制御方策を用いており，正確な自己位置を取得していることを前提としていた [1]．一方，IMU のような自己中心的観測のみに依存する構成の場合は，方位の整合性が崩れやすく，長期的な誘導が困難となる．そこで本研究では，進行方向のアンカーとして，スタート地点から目的地への全体方位角を導入する．この値はエピソード開始時に一意に定義され，学習中は常に一定値として与えられる．

#### 3.2.2　ベクトル観測情報

前節の「全体方位角」を含む一連の幾何・運動情報を，方策学習に用いるベクトル観測情報として設計する．これらは ROS 実機への方策転移を想定し，Unity 上で得られるエージェント基準情報から，IMU やオドメトリを模倣して構成する．各観測量は以下の 5 種類からなる：

(1) **エージェントの相対位置と姿勢**（$\mathbf{p}_{\mathrm{agent}}, \phi_{\mathrm{agent}}$）
エージェントの現在位置と姿勢を，その初期位置との差分として算出したものである．姿勢角は $[-\pi, \pi]$ の範囲とする．このような表現は地図非依存の空間把握に有効とされている [3]．

(2) **目的地の相対位置と姿勢**（$\mathbf{p}_{\mathrm{goal}}, \phi_{\mathrm{goal}}$）
エージェントの初期位置を基準とした目的地との相対位置・姿勢 [3]．姿勢角は $[-\pi, \pi]$ の範囲とする．

(3) **エージェントと目的地の相対方位角と距離**（$\theta, d$）
エージェントから見た目的地の方向とユークリッド距離を表す．$\theta$ は前方を基準とした有向角度であり，ニューラルネットワークへの入力として扱いや

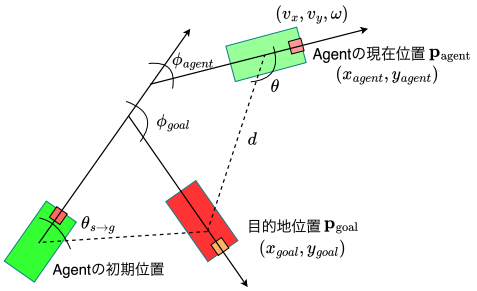



図 5　提案モデルにおけるベクトル観測情報の構成

すくするために $[-1, 1]$ に正規化する．

(4) **並進速度および角速度**（$v_x, v_y, \omega$）
IMU を模倣した動的観測量であり，現在と 1 ステップ前の位置・姿勢の差分から算出する．この情報は環境応答や姿勢変化の理解を助け，方策の安定化に寄与するとされている [6]．

(5) **初期位置から目的地への全体方位角**（$\theta_{s\to g}$）
エピソード開始時におけるスタート地点から目的地への一意な方向角．学習中のすべてのステップで定数として与える．この値も $[-1, 1]$ に正規化する．

ベクトル観測情報の全体構成を概略的に図 5 に示す．

#### 3.2.3　視覚観測情報

第 2.1.1節で述べたエージェント視点カメラから取得した画像に前処理を施し，解像度 $112 \times 84$ のグレースケール画像として CNN ベースの視覚エンコーダに入力することで，空間的特徴量を抽出し，ベクトル観測情報と統合して方策ネットワークの入力とする．

### 3.3　報酬関数の設計

ナビゲーションの安定性と効率性を両立させるために，目標到達時の「成功報酬」，目標へ移動を促す「距離報酬」，および複数の罰則（負の報酬）から成る複合報酬関数を設計する．各報酬の構成を表 1に示す．

「距離報酬」には，Yokoyama らの提案する距離と方向整合性の加重和指標 $J_{\text{nav}} = d + w \cdot \theta$ の変化量 $\Delta J_{\text{nav}}$ を採用した [3]．ここで，$d$ は目的地までのユークリッド距離，$\theta$ は相対角度（ラジアン），重み係数は $w = 0.1$ に設定した．また，$\theta$ の算出は距離が 2 m 未満の場合に限り行い，それ以外では定数 3.14 を与える．これは近距離以外での過剰な整列を抑え，壁衝突を防ぐためである．

さらに，特定の望ましくない行動を抑制し，学習の安定化を目的として，壁との衝突を避けるための「衝突罰則」，行動の遅延を避けるための「遅延罰則」，負の並進速度を抑制する「後退罰則」，速度や方向の過度な振動を抑える「振動罰則」，前後移動方向の頻繁な切替を抑える「切替罰則」を導入した．

表 1　報酬設計

| 報酬項目 | 報酬値 |
|---|---|
| 成功報酬 | +30 |
| 距離報酬 | $\Delta J_{\text{nav}}$ |
| 衝突罰則 | -20 |
| 遅延罰則 (1 ステップごとに) | -0.01 |
| 後退罰則 (負速度時) | -0.01 |
| 振動罰則 (0.2 秒以内移動 0.05 m 未満) | -0.01 |
| 切替罰則 (5 ステップ切替 3 回以上) | -0.01 |

## 4.　シミュレーション環境における検証実験

本章では，訓練過程の各種記録データを通じてモデルの性能と行動特性を分析し，複数の検証用マップでの走行により提案手法の有効性を検証した結果を報告する．

### 4.1　比較モデルの構築方針

提案手法の各設計要素が学習に与える影響を検証するために，図 3 に示した訓練用マップの複数のサブマップ上で，それぞれ 11 体のエージェントを並列に学習させることで，4 種類の比較モデル（A〜D）を構築した．具体的には，「全体方位角の導入」と「LSTM モジュールの有無」という 2 つの要素の組み合わせを変化させて評価した．各モデルに含まれる設計要素の有無を表 2 に示す．

表 2　各モデルにおける設計要素の有無

| モデル | 全体方位角 | LSTM |
|---|---|---|
| A | 有 | 有 |
| B | 無 | 有 |
| C | 有 | 無 |
| D | 無 | 無 |

### 4.2　訓練過程における性能と安定性の比較評価

前節で設計した 4 種類の比較モデル（A〜D）に対して，図 3に示す訓練用マップ上で最大 700 万ステップの強化学習訓練を実施した．図 6は，各モデルを 13 体のエージェントで並列訓練した際の代表的な指標推移を示しており，10,000 ステップごとのエピソードに基づく平均値を可視化している．

この結果から，モデル A は最も高い成功率を安定して示しており，他モデルに対して一貫して優れた性能を達成した．モデル B と C は中程度の性能を維持し，モデル D は成功率が低い傾向を示している．また，モデル A は累積報酬・距離報酬ともに良好である一方，振動・後退などの罰則が最小であり，望ましくない動作の抑制に成功している．さらに，方策エントロピーの推移は，モデル A が過学習を避けつつ行動確定性を高めているのに対し，モデル D は探索不足により早期に収束し多様な戦略の学習に失敗している可能性を示唆している．

以上より，提案手法であるモデル A は，全体的に最も安定かつ高性能なナビゲーション方策を獲得しており，他モデルと比較して明確な優位性が実証された．

### 4.3　テストマップにおける走行性能評価

図 7 に示す 4 種類の検証実験用マップ（T1〜T4）において，モデル A〜D を用いた定量的な走行性能評価実験を実施した．各マップに対して，2.2.2 節で述べた方法により初期位置と目的地を設定し，モデルごとに 100 エピソード（各エピソードは 10,000 ステップ）ずつ走行させた．表 3は，その評価結果として，成功回数・衝突回数・途中終了回数の内訳と，それに基づく成功率を示している．ここで，「成功」とはエージェントが最大ステップ数以内に目的地に到達した場合，「衝突」は障害物に接触してエピソードが強制終了した場合，「途中終了」は最大ステップ数以内に目的地に到達できなかった場合を指す．

T1 は訓練に用いたマップ全体を対象とし，T2 と T3 は左右端の遠距離経路を抽出した構成である．T2 では

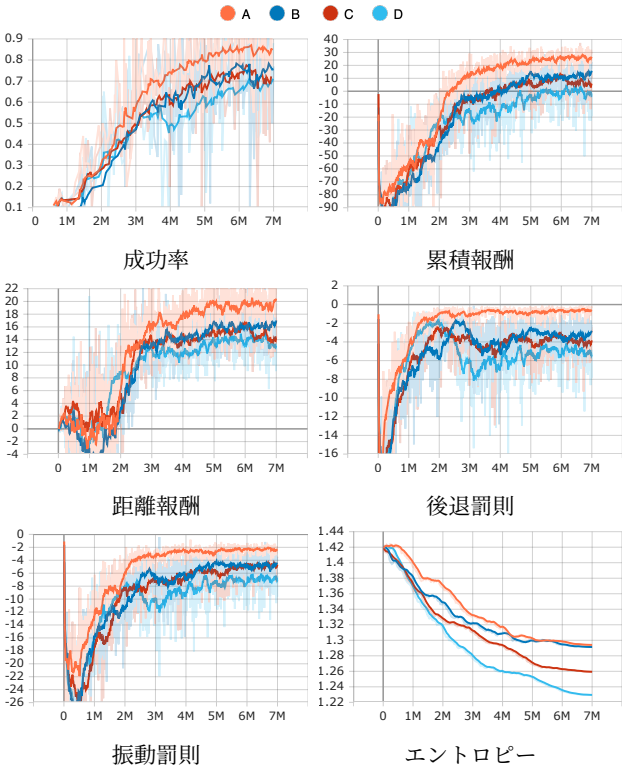



図 6　訓練過程における代表的な報酬と指標の推移

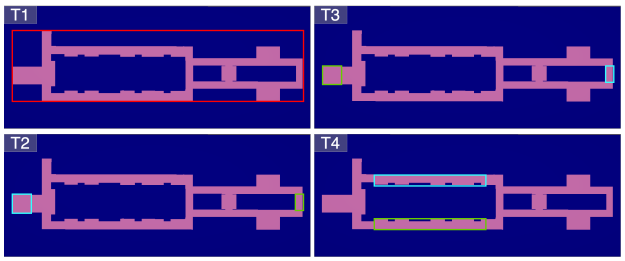



図 7　検証実験用マップ

狭い通路での方向調整，T3 では凹型地形の回避が求められる．T4 は中央に長い廊下区間を含み，遠回り経路を通過しなければ到達できない難易度が高い構造を持つ．

表 3 に示す通り，T1 においてモデル A は成功率 83% を記録し，他モデルを大きく上回った．T2 ではモデル B が 75% とわずかにモデル A（74%）を上回ったが，T3 および T4 のような長距離・複雑経路においては，モデル A がそれぞれ 67%, 59% と最も高い到達率を示し，他モデルに対して明確な優位性を示した．これらの結果から，提案手法に基づくモデル A は，長距離経路や複雑構造を含む環境下の走行で，他の比較モデルに対して明確な優位性を有することが確認された．

## 5.　まとめと今後の課題

本稿では，任意の目的地への到達を目指す強化学習ナビゲーションにおいて，視覚および IMU を想定した観測設計を行い，安定した方策学習を実現するモデルを構築した．また，複数の比較モデルとの走行性能評価により，提案手法に基づくモデル A が最も高い成功率を示し，特に長距離や複雑な経路条件での高い到達性能を確認した．

今後の課題としては，まず実機 ROS ロボットへの展開を見据えた Sim2Real 転移の検証が挙げられる．現時点ではシミュレーション環境での検証に留まっており，

表 3　各モデルの走行性能および失敗要因の比較

| マップ | モデル | 成功回数 | 衝突回数 | 途中終了回数 | 成功率 |
|---|---|---|---|---|---|
| T1 | A | 83 | 13 | 4 | 83% |
| | B | 48 | 32 | 21 | 48% |
| | C | 37 | 35 | 28 | 37% |
| | D | 43 | 29 | 28 | 43% |
| T2 | A | 74 | 25 | 1 | 74% |
| | B | 75 | 24 | 1 | 75% |
| | C | 0 | 73 | 27 | 0% |
| | D | 20 | 23 | 57 | 20% |
| T3 | A | 67 | 32 | 1 | 67% |
| | B | 4 | 51 | 45 | 4% |
| | C | 13 | 85 | 2 | 13% |
| | D | 0 | 13 | 87 | 0% |
| T4 | A | 59 | 32 | 9 | 59% |
| | B | 8 | 46 | 46 | 8% |
| | C | 7 | 26 | 67 | 7% |
| | D | 0 | 13 | 87 | 0% |

現実環境におけるセンサーノイズや認識誤差への耐性を含めた評価が必要である．また，モデル A の挙動を詳細に分析したところ，右方向への旋回がやや多く発生する傾向が観察された．このような片方向への行動偏りは，強化学習においてしばしば報告される局所的な最適解への収束の一形態であり，本研究においては訓練時に用いたマップ分割の都合により，右折経路の方が目的地に到達しやすいケースが多く存在したことが原因と考えられる．今後は，地図の構造や初期位置と目的地の分布に対するバイアスを抑制し，よりバランスの取れた経路探索能力を実現するための学習環境の改良が求められる．

## 参　考　文　献

[1] R. Tsuruta and K. Morioka, “Autonomous navigation of a mobile robot with a monocular camera using deep reinforcement learning and semantic image segmentation,” in *IEEE/SICE International Symposium on System Integration (SII)*, pp. 1107–1112, IEEE, 2024.

[2] 土屋一朗，森岡一幸，“Rgb-d カメラから得られる情報を用いた深層強化学習に基づく自律移動ロボット，” in 第 *21* 回計測自動制御学会システムインテグレーション部門講演会 *(SI2020)*, no. 1H3-03, 2020.

[3] N. Yokoyama, A. Clegg, J. Truong, E. Undersander, T.-Y. Yang, S. Arnaud, S. Ha, D. Batra, and A. Rai, “Asc: Adaptive skill coordination for robotic mobile manipulation,” *IEEE Robotics and Automation Letters*, 2023.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” in *arXiv:1707.06347*, 2017.

[5] M. Hausknecht and P. Stone, “Deep recurrent q-learning for partially observable mdps,” in *Proceedings of the AAAI Fall Symposium Series*, 2015.

[6] Y. Chen, R. Yang, Y. Zhang, Q. Tian, G. Li, L. Sun, and X. Huang, “Inertial navigation meets deep learning: A survey of current trends and future directions,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14509–14528, 2022.