Map-Free 強化学習ナビゲーションの Sim2Real による実環境での走行

○天野 大輔（明治大学），森岡 一幸（明治大学）

Sim2Real Navigation Based on Map-Free Reinforcement Learning in Real Environments

○ Daisuke AMANO (Meiji University), and Kazuyuki MORIOKA (Meiji University)

Abstract: This study investigates Map-Free reinforcement learning (RL) navigation without relying on 2D LiDAR or pre-built maps. An RL policy trained in a Unity simulation was transferred to a real robot within a Sim2Real framework. In real environments, binary ground segmentation was achieved using a Unet-ResNet50 model trained with masks generated by SAM2. For self-localization, Visual-Inertial Odometry (VIO) was employed using a stereo camera and IMU. Real-world experiments demonstrated the feasibility and effectiveness of the proposed approach.

1. 緒言

近年，移動ロボットの自律ナビゲーションにおいて，深層強化学習を用いた方策学習が注目されている．このアプローチは，センサー情報を直接入力とし，環境との相互作用を通じて目的地までの行動モデルを獲得できる点に利点がある．従来研究では，2D LiDAR や深度カメラを用いた事例が多く報告されているが[1), 2)]，これらは高価なセンサーや事前の地図作成に依存することが課題である．一方，カメラ画像を入力とする手法は低コストかつ柔軟であるが，実環境では照明条件や外乱の影響を受けやすく，シミュレーション環境からの転移が困難であると指摘されている[1)]．

著者らの以前の研究[3)]では，Unity 上のシミュレーション環境において，視覚観測とベクトル観測を統合した Map-Free 強化学習ナビゲーション方策を提案し，任意の目的地への到達を実現可能であることを示した．しかし，その成果はシミュレーション内にとどまっており，実機環境に適用した検証は十分に行われていなかった．

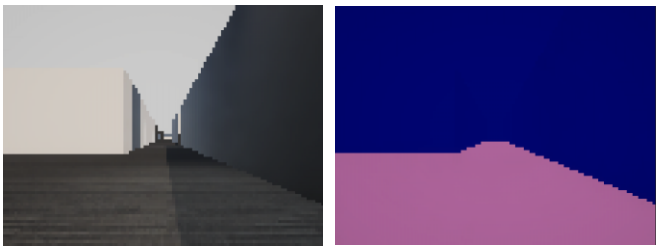
そこで本研究では，実環境における Sim2Real 転移を目的とし，実機走行実験を通じて Map-Free 強化学習ナビゲーションの有効性を検証する．具体的には，セグメンテーションモデルによる地面領域の二値化と，ステレオカメラ ZED2i が提供する Visual-Inertial Odometry (VIO) に基づく自己位置推定を組み合わせ，シミュレーションで獲得した方策を実機ロボットに適用する．実験の結果，事前の地図作成や LiDAR に依存せずに，安定した自律走行を実現可能であることを確認した．

2. シミュレーション環境を用いたナビゲーション方策の獲得

本研究で用いるナビゲーション方策は，著者の以前の研究[3)]において報告した視覚およびベクトル観測情報を統合した強化学習モデルを基盤としている．本章では，そのモデル構造，シミュレーション環境での二値化表現，観測設計，および訓練の概要について述べる．

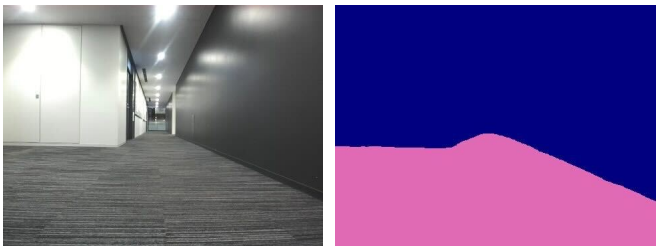
2.1 強化学習モデルの構造

本研究では強化学習アルゴリズムとして Proximal Policy Optimization (PPO)[4)] を採用し，さらに観測系列の時間的依存関係を捉えるために LSTM を組み込んだ方策モデルを設計した．これにより，環境ノイズや短期的な状態変動に対しても安定したナビゲーション行動を獲得できることを期待した．



(a) Original (Sim) (b) Binarized (Sim)

Unity simulation (Sim) environment



(c) Original (Real) (d) Binarized (Real)

Real environment

Fig. 1 The view from the agent's perspective

2.2 シミュレーション環境と地図の二値化

強化学習においては，色情報や照明条件の揺らぎが学習の不安定要因となる．この影響を抑えるため，床面を赤紫色，壁面を暗青色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)]．**Fig. 1(a)** に示すように，Unity シミュレーション環境におけるエージェント視点の元画像から，**Fig. 1(b)** のような二値化画像を得ることで，視覚センサー入力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．

2.3 観測情報

本研究で用いる観測情報は，二値化した視覚観測と，**Fig. 2** に示すベクトル観測を統合したものである．視覚観測としては，前述の二値化画像を解像度 112×84 のグレースケール画像に変換して CNN エンコーダに入力した．さらに，ベクトル観測として，エージェントの相対位置と姿勢，目的地との相対位置と姿勢，距離

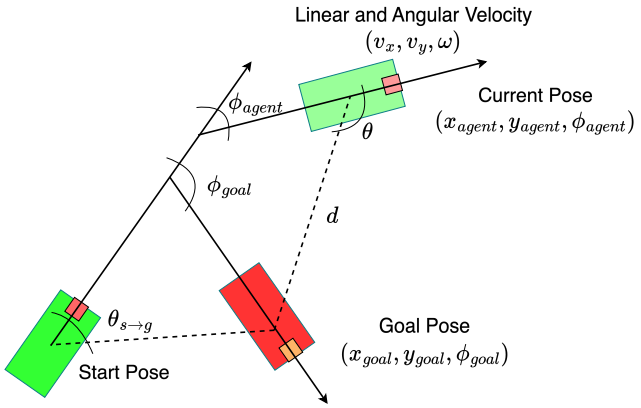


Fig. 2 Vector observations used in the RL model

と方向角，IMU を模した線速度および角速度情報，および初期位置から目的地への全体方位角を設計に含めた．これにより，視覚入力だけでは不安定となる距離推定や方向認識を補完し，実機 ROS 環境への転移を考慮した観測設計とした．

2.4 強化学習モデル訓練の概要

ナビゲーションの安定性と効率性を両立させるために，本研究では **Table 1** に示す複合報酬関数を設計した．まず，目標到達時には「成功報酬」を与え，タスク達成を直接的に強化する．さらに，目標への接近を促すために「距離報酬」を導入した．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 ΔJ_{nav} を用いた．ここで，d は目的地までのユークリッド距離，θ は相対角度であり，重み係数 w は 0.1 に設定した．また，θ の算出は距離が 2 m 未満の場合に限定し，それ以外では定数 3.14 を与えることで，近距離以外での過剰な整列を抑え，壁衝突を防ぐようにした．

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複数の罰則を導入した．「衝突罰則」は障害物との接触時に与え，「遅延罰則」は各ステップごとに小さな負の値を与えることで停滞を防ぐ．また，「後退罰則」は負の並進速度を抑制するために導入し，「振動罰則」は短時間に微小な揺れを繰り返す不安定な行動を抑える．さらに，「切替罰則」は 5 ステップ以内に 3 回以上の前後切替が生じた場合に課し，進行方向の頻繁な反転を抑制する．

これらの報酬設計に基づき，Unity シミュレーション環境において 70 万ステップの強化学習を行い，ナビゲーションモデルを獲得した．

3. 実環境への転移手法

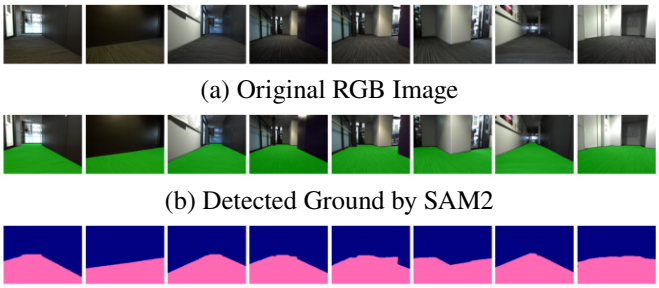
シミュレーション環境で獲得した方策を実機に適用するためには，観測表現の違いやセンサーノイズなどによる Sim2Real ギャップを補う必要がある．本章では，実環境での観測情報を整備する手法として，まず地面領域を抽出する二値化分割モデルの構築について述べ，続いてステレオカメラと IMU を用いたセンサー融合による自己位置推定について説明する．

3.1 地面二値化分割モデル

シミュレーション環境では，地面と障害物を二値化することで観測の単純化し，強化学習の安定した方策獲得につなげてきた．しかし実環境では RGB 画像そのままでは同様の表現が得られず，シミュレーションと

Table 1 Reward settings

Reward Type	Reward Value
Success Reward	+30
Distance Reward	ΔJ_{nav}
Collision Penalty	-20
Slack Penalty (per step)	-0.01
MoveBack Penalty (negative linear velocity)	-0.01
Oscillatory Penalty (movement <0.05 m within 0.2 s)	-0.01
Switch Penalty (≥ 3 switches within 5 steps)	-0.01



(a) Original RGB Image

(b) Detected Ground by SAM2

(c) Binarized Mask

Fig. 3 Examples of the constructed dataset

のギャップが生じる．そこで本研究では，実環境においても可走行領域を明示的に区別し，Unity シミュレーション環境で用いた二値化表現に近い観測を得るため，地面領域を抽出する二値化分割モデルを構築した．

3.1.1 データセット構築

シミュレーション環境で用いた二値化表現を実環境でも再現するためには，まず高精度な教師ラベルを備えたデータセットが必要となる．従来の平面検出によるマスク生成[1] ではラベル精度に課題があり，そのままでは学習の安定性に影響する可能性がある．

本研究では，**Fig. 3(a)** に示すように著者らが所属する大学の廊下において ZED2i ステレオカメラの左眼で撮影した 191 枚の画像を用いた．これらに対して Meta 社が公開する高精度セグメンテーションモデル SAM2[5] を適用し，**Fig. 3(b)** に示すように地面領域を検出した．さらに検出結果を基に，**Fig. 3(c)** のような二値化マスクを生成した．

最終的に，Fig. 3(a) と Fig. 3(c) を組み合わせて，訓練に用いるデータセットを構築した．

3.1.2 二値化分割モデルの獲得と評価

SAM2 は高精度である一方，計算資源を多く必要とし実機上でのリアルタイム処理には不向きである．そこで本研究では，SAM2 で生成したマスクを教師データとする上述のデータセットを用い，より軽量な Unet-ResNet50 モデルの学習を行った．データセットは 70%を学習用，15%を検証用，15%を評価用にランダムに分割し，学習データには一般的なデータ拡張を施して汎化性能の向上を図った．損失関数には DiceLoss を，最適化手法には Adam を採用し，80 epoch の学習を行った．また各 epoch で検証データに対して IoU を算

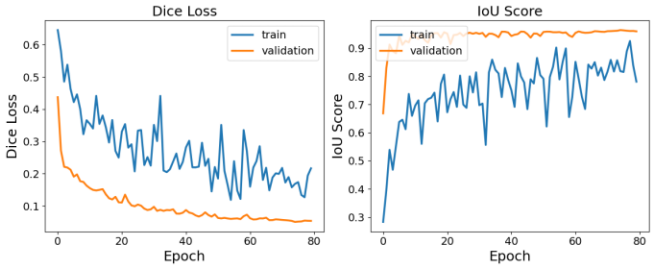


Fig. 4 Learning curves of the segmentation model

(a) Original (b) Detected (c) Binarized

Fig. 5 Detected obstacles in the real environment

出し，最高スコアを記録したモデルを保存した．

学習過程における指標推移を **Fig. 4** に示す．Dice Loss と IoU の変化から，モデルが安定して学習していることが確認できる．さらに実環境での推論例として，**Fig. 1(c)** に入力 RGB 画像を，**Fig. 1(d)** に推論で得られた二値化画像を示す．加えて，現実環境において障害物が存在する場合でも本モデルは検出可能であることを確認しており，その例を **Fig. 5** に示す．

3.2 多センサー融合による自己位置推定

実環境における自己位置推定には，ZED2i ステレオカメラが提供する Visual-Inertial Odometry (VIO) を用いた．この VIO は，ステレオカメラの視覚オドメトリと内蔵 IMU（加速度計・ジャイロ・磁力計）の情報を組み合わせて得られるものであり，純粋な視覚オドメトリと比較して長時間走行におけるドリフトが少なく，より安定した推定が可能となる．結果として得られるオドメトリ座標系における姿勢推定は，純粋な視覚オドメトリと比較して長時間の走行でもドリフトが少なく，より安定した位置推定を可能にする．

実環境では，強化学習モデルの観測ベクトルを対応させるため，初期位置を原点として再定義し，エージェントの相対的な位置・姿勢・速度を計算した．これにより，シミュレーション環境で設計したベクトル観測と整合性を保つことができる．最終的に，Fig. 2 に示すベクトル観測情報を構成した．

以下に，ベクトル観測の五つの情報源を列挙する．

1) **ロボットの相対位置と姿勢**：ZED2i の VIO により得られる odom 座標系に基づき，初期位置の座標を基準として $(x_{robot}, y_{robot}, \phi_{robot})$ を算出．
2) **目的地の相対位置と姿勢**：目的地の位置 $(x_{goal}, y_{goal}, \phi_{goal})$ を同じく初期位置の座標を基準として相対表現に変換．
3) **ロボットと目的地の相対方向と距離**：ロボットから見た目的地方向の角度 θ と距離 d を算出．
4) **速度情報**：直前の位置差分から得られる (v_x, v_y) に加え，IMU のジャイロ出力から角速度 ω を算出．
5) **初期位置から目的地への全体方位角**：初期位置位置から目的地方向への角度 $\theta_{s\to g}$ を算出．

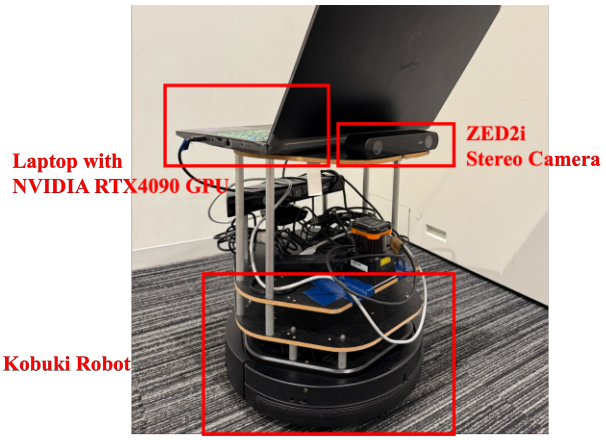


Fig. 6 Experimental hardware configuration

Table 2 System configuration of the experimental system

Component	Specification / Purpose
Mobile Robot Base	Kobuki (base platform)
Stereo Camera	ZED2i (RGB input, VIO)
Computing Unit	NVIDIA RTX4060 Laptop
Software Environment	Ubuntu 24.04, ROS2 Jazzy

4. 実機走行検証実験

本章では，前章で述べた実環境への転移手法を組み合わせることにより，移動ロボットを用いた実機走行実験を通じて評価を行い，提案手法である Map-Free 強化学習ナビゲーションの有効性を検証する．

4.1 システム構成

実機走行実験には，移動ロボットプラットフォームとして Kobuki を用い，上部にステレオカメラ ZED2i を搭載して環境画像および自己位置推定に必要な VIO 情報を取得した．計算処理には，GPU（NVIDIA RTX4060 Laptop）を搭載したノート PC を用い，ROS2 環境上で二値化分割モデルと強化学習に基づくナビゲーション方策を実行した．実験システムのハードウエア構成を **Fig. 6** に，主要な仕様を **Table 2** に示す．また，実験中の ROS2 ノードグラフを **Fig. 7** に示す．

4.2 実験方法

実機走行実験は，所属する大学の廊下で行った．**Fig. 8** に示すように，エージェントの初期位置 (Start) から目的地 (Goal) までの直線距離は約 50 m である．

まず，リモコンを用いてロボットを初期位置から目的地まで走行させ，その際に VIO により推定された目的地の座標を記録した．その後，ロボットを再び初期位置に戻し，システムを初期化した，初期位置から目的地までの約 53 m の走行（緑色の経路）と，目的地から初期位置までの約 57 m の走行（赤色の経路）を連続して行い，合計約 110 m の長距離走行実験を実施した．

4.3 実験結果と考察

実験の結果，初期位置から目的地までの走行，および目的地から初期位置までの走行のいずれにおいても，ロボットは比較的短時間で安定して到達することが確認された．それぞれの走行時間は約 150 秒および約 170 秒であり，その往復の走行軌跡を **Fig. 9** に示す．特

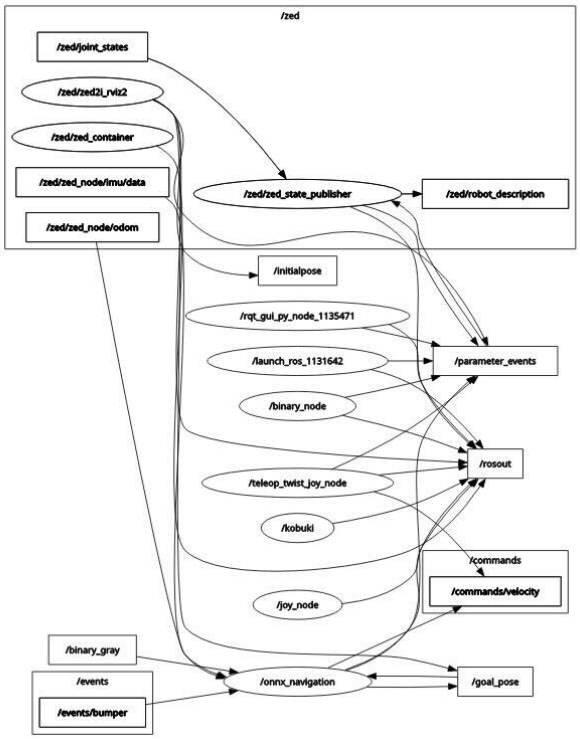


Fig. 7 ROS2 node graph during the experiment

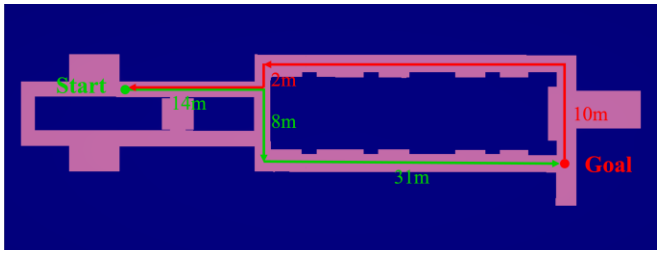


Fig. 8 Experimental environment map

に直線区間における走行は非常に安定していた．また，曲がり角において壁面に接近する場面が見られたが，ロボットは角度を修正して走行を継続し，衝突することなく通過することができた．これにより，本研究で得られた強化学習ナビゲーションモデルが高い安定性を有することが示された．

一方で，目的地から初期位置に戻る際には，ロボットはおおむね安定して初期位置付近に到達したものの，最終的な自己位置推定においては実際の初期位置との間に約 2 m の誤差が残った．この要因としては，走行に伴う VIO の累積誤差が影響していると考えられる．さらに，得られた走行データと，走行環境のレイアウトから想定される理想的な経路を比較し，累積的な RMSE の推移を **Fig. 10** に示す．この図より，全体の RMSE は最終的に約 2.84 m に収束していることが確認され，自己位置推定の精度に一定の課題が残ることが明らかとなった．

以上の結果から，シミュレーション環境において学習した強化学習ナビゲーションモデルが，実環境においても安定して動作可能であることが確認された．特に，本実験では LiDAR を用いず視覚情報のみに依存した構成であっても，かつ事前に地図を作成することなく，ロボットは目的地までの安定したナビゲーションを実現できることが示された．

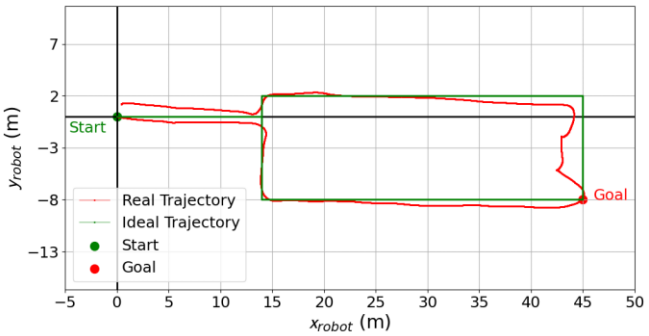


Fig. 9 Trajectory of the robot during the experiment

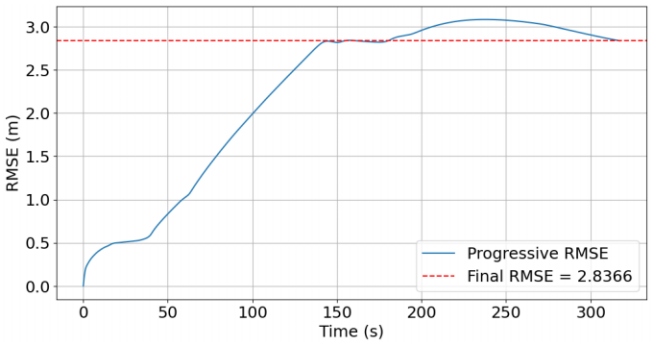


Fig. 10 Progressive RMSE of trajectory vs ground truth

5. 結言

本研究では，シミュレーション環境において獲得した Map-Free 強化学習ナビゲーションモデルと，その Sim2Real 転移手法を実機環境に適用し，実環境での走行実験を通じて検証を行った．その結果，提案手法は事前に地図を作成せず，また LiDAR を用いない視覚情報に基づく構成であっても，目的地までの安定したナビゲーションを実現できることが示された．

一方で，長距離走行においてはオドメトリの累積誤差に起因する位置推定のドリフトが残り，最終位置において数メートル規模の誤差が生じることが確認された．今後は，この課題を解決するために，回環検出を備えた SLAM 手法を導入し，長距離走行時の全体的な自己位置推定の精度向上を目指す．

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, IEEE/SICE International Symposium on System Integration (SII), IEEE (2024), pp. 1107–1112, DOI: 10.1109/SII58957.2024.10417188.

[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，計測自動制御学会 SI 部門講演会 SICE-SI 予稿集 (2020).

[3] 天野大輔，森岡一幸: 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，日本ロボット学会学術講演会 (2025).

[4] J. Schulman et al.: Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).

[5] N. Ravi et al.: SAM 2: Segment Anything in Images and Videos, arXiv preprint arXiv:2408.00714 (2024).