

# 不整地・障害物フィールドに対応する 4 脚ロボット制御のための強化学習手法の実装と比較

村上研究室 f21040 5J13 小山田智典

## 背景

近年、脚式ロボット (e.g. Unitree Go2, ANYmal, Unitree G1) の市販流通が進んでおり、畑や山などの不整地環境での活用が期待されている。特に 4 脚ロボットはその安定性から実環境での試験運用が進んでいる。しかし、不整地歩行のための強化学習モデルは学習済みのモデルウェイトが公開されておらず、ほとんどの場合、独自に訓練する必要がある。そこで、4 脚ロボットの不整地歩行を可能にするソフトウェア・ハードウェアの基盤を構築することを目的とし、既存手法の調査と再実装、機能の比較を行う。

## 実装

検証のために、モデルおよび強化学習のアルゴリズムとシミュレーション環境を実装した。

### 実装に使用したライブラリ

- PyTorch: 深層学習フレームワーク
- Genesis: ロボットシミュレーション環境構築ライブラリ

### 強化学習環境の実装

- ロボットモデル: Unitree Go2 (4 脚ロボット)
- フィールド: 平坦な地面, 不整地フィールド
- 観測情報: 60 次元の状態情報
  - ボディの速度  $v_{xyz}$  (ボディ座標系)
  - ボディの回転角速度  $\omega_{xyz}$  (ボディ座標系)
  - 関節の角度  $\theta$ , 角速度  $\theta'$ , トルク  $\tau$
  - 一つ前の行動  $a_{t-1}$
  - 目標動作コマンド  $c$  (前進速度, 横移動速度, 回転速度)
- 報酬設計: 12 項目
  - xy 速度, yaw 角速度: 目標動作コマンドとの一致度
  - 足の先端位置: 移動軌跡との近さ
  - 足の接地状態: 対角線の足が接地しているか
  - 肩の高さ: 地面と一定の距離を保っているか
  - ボディの安定性: 重力ベクトルの z 成分の大きさ
  - 行動の最小化: 初期姿勢からの行動変化の抑制
  - 縦揺れ抑制, 回転抑制: ボディの不必要な振動と回転を抑制する
  - 足先, 関節の円滑性: 足の先端と関節角度の変化の滑らかさ
  - 衝突回避: ボディと地面の衝突を避ける

## 比較・考察

Table 1 に示す 3 つの手法を実装し、歩行性能の比較を行った。

Table 1: 平地・不整地歩行の可否検証結果

手法	平地歩行	不整地歩行
PPO + MLP	○	×
PPO + RNN	○	×
SLR	○	○

MLP では過去の情報を用いず、RNN では効率的な過去情報の伝搬を学習することができなかったため、不整地歩行は困難であった。一方で SLR は、過去の観測情報を時系列的に一貫性のある潜在表現に変換して、効率的に利用することで路面状況を推測することができ、不整地歩行が可能になったと考えられる。

## 今後の展望

現状の手法では、RGB カメラや Depth カメラなどの視覚情報を利用していないため、障害物の認識が困難であり複雑な不整地歩行には対応できていない。

今後は、より走破性の高い方策モデルの構築を目指し、視覚情報を活用した手法 (e.g. [1]) の実装を行い、構築した方策モデルを用いて、実機ロボットへの移植や機能追加 (e.g 障害物回避, 音声制御) を行う予定である。

## 強化学習手法: Proximal Policy Optimization

Proximal Policy Optimization (以下, PPO) は Actor-Critic 法をベースとする強化学習であり、ロボット制御に限らず様々な分野の強化学習タスクで高い性能を発揮している。PPO は、方策関数 (Policy) の勾配を安定的に更新することを目的としており、クリッピング手法を用いることで急激な方策更新による性能の劣化を防いでいる。PPO の最大化目的は以下の式で表される。(出典: [2])

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \quad r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}$$

where  $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[ \min \left( r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

## 先行研究: Self-learning Latent Representation

Self-learning Latent Representation (以下, SLR) は PPO をベースとした手法であり、過去の観測情報を埋め込んで MLP に入力することで、不整地歩行に必要な情報を効率的に学習することを目的としている。

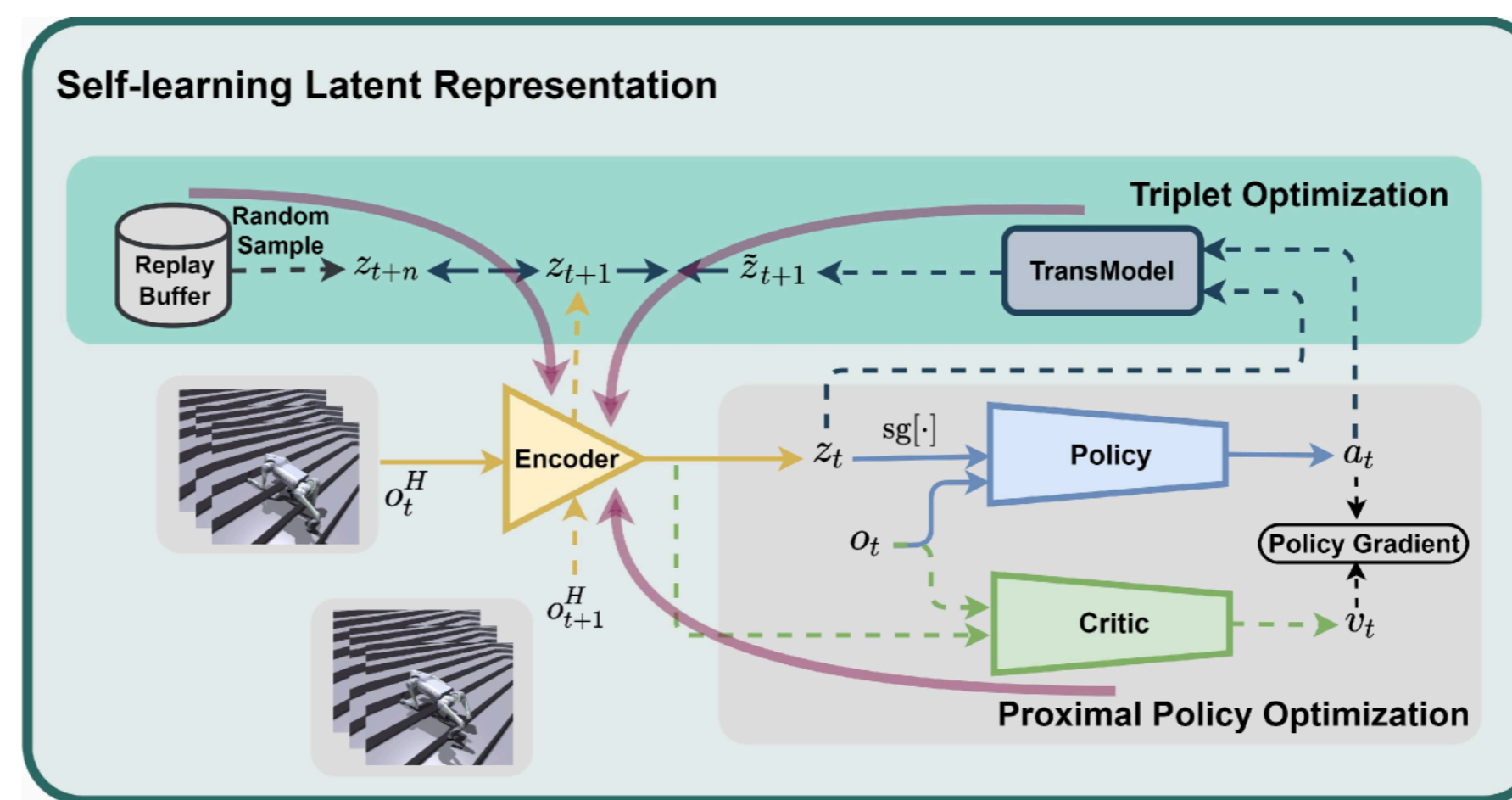


Figure 1: SLR の学習フレームワーク (出典: [3])

### SLR が導入するモデル

- Encoder: 過去の観測情報を低次元の潜在表現に変換するモデル

$$z_t = \varphi(o_t^H) \quad \text{※ } o_t^H \text{ は } \{t-h:t\} \text{ の範囲の観測情報}$$

- TransModel: 潜在表現の時系列変化を予測するモデル

$$\tilde{z}_{t+1} = \mu(z_t, a_t)$$

SLR の損失関数は、PPO の損失関数に加えて、潜在表現  $z$  の時系列整合性を保つためのトリプレット損失を導入している。

$$\mathcal{L}_{\text{trip}} = \max(\|z_{t+1} - \tilde{z}_{t+1}\| - \|z_{t+1} - z_{t+n}\| + m, 0), \quad s.t. \quad n \neq 1$$

## 参考文献

- [1] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, “Learning robust perceptive locomotion for quadrupedal robots in the wild,” Science Robotics, vol. 7, no. 62, Jan. 2022, doi: [10.1126/scirobotics.abk2822](https://doi.org/10.1126/scirobotics.abk2822).
- [2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms.” [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [3] S. Chen et al., “SLR: Learning Quadruped Locomotion without Privileged Information.” [Online]. Available: <https://arxiv.org/abs/2406.04835>