F-020

行動変換モデルを導入した Sim-to-Real 強化学習による油圧ショベルの自動操作の検討

Automated Operation of Hydraulic Excavators Using Sim-to-Real Reinforcement Learning with Action Transformation

中西 亮太[1)] 青木 大地[1)] 西原 賢太[1)] 清水 彰馬[1)]
Ryota Nakanishi Daichi Aoki Kenta Nishihara Shoma Shimizu

許 凱宇[2)] 小松 琢也[2)] 内田 絢斗[1)] 白川 真一[1)]
Kaiyu Xu Takuya Komatsu Kento Uchida Shinichi Shirakawa

1 はじめに

建設機械の操作には熟練者の高度な技術が必要であり，操作技術の習得には長期間の訓練が必要である．特に，熟練運転手が実際に建設機械を操作して得られたデータをもとに，動作速度や制御パラメータを調整するチューニング作業は，熟練技術者による操作が不可欠な行程の一つである．例えば，油圧ショベルのチューニング作業では，バケットの刃先を一定の方向で引くすきとり操作などが熟練運転手により行われている．しかし，建設業界では熟練運転手の高齢化とそれに伴う人材不足が深刻な課題となっており，熟練技術が必要な操作を持続的に行うことが難しくなると懸念されている．令和 5 年の国土交通省の調査 [1] によれば，建設機械の操作に熟練技術を要する作業が増加している一方で，それを担う人材の不足が報告されている．

このような状況から，建設機械の自動操作を実現する技術が注目されている．直接的なアプローチとして実際の建設機械を利用した強化学習が考えられるが，実機の稼働に時間がかかることや，学習中に危険な操作や機器の破損が発生する可能性があり，現実的ではない．そこでシミュレーション環境を用いて自動操作モデルを構築し，運用環境に転用する Sim-to-Real Transfer [2] が検討されている．しかしシミュレーション環境で再現された建設機械は運用環境と異なる振る舞いをする可能性があり，この環境の差異による性能低下が課題となっている．

そこで本研究では，Sim-to-Real 強化学習手法である Reinforced Grounded Action Transformation (RGAT) [3] に基づき，環境の差異に頑健な油圧ショベルの自動操作モデルの獲得を目指す．RGAT は学習環境と運用環境の差異を吸収する行動変換モデルを導入し，学習環境上での強化学習を通じて運用環境に適用できる自動操作モデルを学習する強化学習手法である．RGAT では学習中に運用環境で方策を動かし，学習用の軌道を獲得する必要があり，本研究での問題設定ではそのまま利用することが難しい．そこで本研究では，事前に運用環境で収集した熟練運転手の操作軌道のみを利用し，学習中には運用環境でのデータ収集を行わないよう RGAT のアルゴリズムを変更する．そしてチューニング作業として行われる空中すきとり操作を対象に，低精度シミュレータで学習した自動操作モデルを高精度シミュレータ上で評価することで，提案手法で獲得した自動操作モデルの性能を評価する．

1） 横浜国立大学. Yokohama National University.
2） 株式会社小松製作所. Komatsu Ltd.

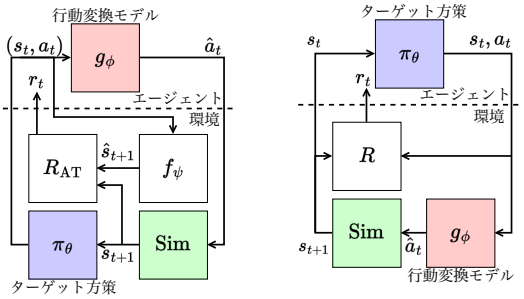

図 1 RGAT における行動変換モデルの学習（左図）とターゲット方策の学習（右図）

2 関連研究

2.1 Reinforced Grounded Action Transformation

Reinforced Grounded Action Transformation (RGAT) は Sim-to-Real Transfer を実現する強化学習手法であり，行動変換モデルを導入することで，運用環境下で適切に動作する方策の学習環境下での獲得を可能にする．行動変換モデルは，入力として状態[1)]と運用環境に適した行動を受け取り，出力として学習環境に適した行動を返す．このモデルにより，学習時に方策が決定した「運用環境に適した行動」が「学習環境に適した行動」に変換されて学習環境に渡される．こうした仕組みによって，運用環境下で適切に動作する方策を学習環境下で学習できるようになる．

RGAT における学習では，図 1 に示す行動変換モデル g_ϕ の学習とターゲット方策 π_θ の学習を繰り返す．行動変換モデルの学習では，状態 s_t と行動 a_t から運用環境における次状態を予測する順方向遷移モデル f_ψ を導入する．この順方向遷移モデルは，現在のターゲット方策を利用して運用環境から収集した実軌道 $\boldsymbol{\tau}_{\mathrm{real}}$ を利用した教師あり学習により獲得する．そして，行動変換モデルにより変換された行動 $\hat{a}_t$ をシミュレータに与えた際に得られる次状態と順方向遷移モデルの予測値との予測誤差を負の報酬として，行動変換モデルを強化学習により更新する．これにより，シミュレータの出力が運用環境での次状態に近づくよう行動を変換する行動変換モデルの獲得を目指す．またターゲット方策の更新では，行動変換モデルとシミュレータを通じて次状態を計算する環境を構築し，強化学習によりターゲット方策を更新する．これにより，運用環境上で高い性能を保つターゲット方策の学習が可能となる．この一連のプロセスをターゲット方策が運用環境で一定以上の改善を示すまで繰り

1） RGAT は状態空間は運用環境と学習環境で同じであると仮定する．

アルゴリズム 1 Reinforced Grounded Action Transformation

Input: ターゲット方策 π_θ，行動変換モデル g_ϕ，順方向遷移モデル f_ψ，各モデルの初期パラメータ θ, ϕ, ψ，方策改善手法 `optimize1`，`optimize2`

1: **while** ターゲット方策が運用環境で改善される **do**
2: ターゲット方策 π_θ を用いて運用環境上での軌道 $\tau_{\text{real}} \leftarrow \{((s_0, a_0), s_1), ((s_1, a_1), s_2), \dots\}_{\text{real}}$ を収集.
3: 軌道 τ_{real} を用いて順方向遷移モデル f_ψ を学習.
4: 報酬 $r_t^{\text{AT}} = -\|f_\psi(s_t, a_t) - s_{t+1}\|^2$ に基づき，学習環境で `optimize1` により行動変換モデル g_ϕ を更新.
5: タスクにおける報酬 R を用いて，学習環境で `optimize2` によりターゲット方策 π_θ を更新.
6: **end while**

返すことで，運用環境で適切に動作するターゲット方策の獲得を目指す．RGAT の更新手順をアルゴリズム 1 に示す．

2.2 強化学習を用いた建設機械の自動運転

建設機械の自動運転に関して強化学習を活用したアプローチが研究されている [4, 5, 6]．特に，文献 [6] では，本研究と同様の問題設定である油圧ショベルの空中すきとり操作の自動操作モデルの開発を行っている．この研究では，まず熟練運転手の操作データを用いて自動操作モデルの教師あり学習を行い，その後に仮想環境上で強化学習を行う 2 段階の学習方法を検討している．この学習方法で獲得された自動操作モデルは学習環境においては十分な性能を発揮するが，運用環境上では性能低下によりタスクの要求基準を満たせないことが課題である．また，強化学習後に改めて教師あり学習を行うことで性能低下の抑制を実現しているが，熟練運転手を代替できるような高性能な自動操作モデルは得られていない．

3 本研究が対象とする油圧ショベル

本節では，本研究の対象である油圧ショベルの特徴，取得可能なセンサデータ，およびセンサデータをもとに作成した車体状態について説明する．

3.1 油圧ショベルの特徴

油圧ショベルの構造を図 2 に示す．油圧ショベルは，運転席の前に配置された作業機と呼ばれる腕状の機構を動かすことで作業を行う．作業機は，先端の「バケット」，バケットと連結する「アーム」，さらにアームと連結する「ブーム」から構成される．これらは油圧シリンダによって接続されている．エンジンの動力は油圧力に変換され、その力で油圧シリンダの伸縮が制御される．

運転手は 2 本のレバーを操作し，油圧シリンダに加わる油圧力を調整して各部位を動かす．本研究では，運用環境上で運転手が行うこれら 2 本のレバー操作によって得られた操作データをもとに学習する．

3.2 油圧ショベルが行う作業

建設現場において，油圧ショベルは掘削作業，積込作業およびすきとり作業などを行う．本研究ではすきとり作業を模擬した空中操作を対象とする．すきとり作業は，基準面に沿ってバケットを奥から手前へ水平に動かし，地面の起伏を削る作業である．本研究では特に，建設機械の操作性のチューニングとエネルギー消費量計測の工程で必要とされる，土に接触しない空中すきとり操作を扱う．図 3 に示すように，空中すきとり操作においては，基準面（図中の点線）に沿ってバケットを正確に動かす必要があり，繊細な操作技術が求められる．

図 2 油圧ショベルの構造

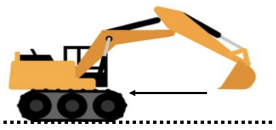
図 3 空中すきとり操作

3.3 操作データについて

油圧ショベルから取得できる操作データは，運用環境の場合は熟練運転手が空中ですきとり操作を行うことで収集された時系列データである．学習に使用したデータのサンプリング周期は 0.05 秒である．各時刻のデータは，油圧ショベルの状態を示すセンサデータと，それに対する熟練運転手のレバー操作量を表す操縦データからなる．

3.3.1 操縦データ

本研究では，同一の運転手が作業機以外を固定した状態で連続して空中ですきとり操作を行ったデータを使用する．運転席には 2 本のレバーがあり，これらを前後に倒すことで油圧シリンダを操作する．レバーの動きを-100 から 100 の範囲で表した値を操縦データとして取得する．これをレバーレートと呼ぶ．レバーレートが正値の場合はシリンダが伸び，負値の場合は縮む．

3.3.2 センサデータ

油圧ショベルに取り付けられたセンサから得られるセンサデータは次に示すとおりである．

- (ブーム/アーム) シリンダ速度 [mm/s]
- (ブーム/アーム) シリンダ長 [mm]
- (フロント/リア) ポンプ圧 [MPa][2)]
- (ブーム/アーム) ボトム圧 [MPa][3)]
- (ブーム/アーム) ヘッド圧 [MPa][4)]
- (フロント/リア) ポンプ容量 [cm^3/rev]
- エンジン回転数 [rpm]
- エンジントルク [N·m]
- (ブーム/アーム/バケット) シリンダ有効力 [kN]

3.3.3 センサデータから作成した車体状態

本研究で行動変換モデルの学習およびターゲット方策の事前学習に使用する操作量と状態量を表 1 に示す．上側 2 行が操作量，以降の 7 行が油圧ショベルの取り付けたセンサから直接得られる状態量である．さらに，油圧ショベルの動作特性をより正確に捉えるため，表の下側 11 行に示す 11 種類の車体状態をセンサデータをもとに作成した．

フロントポンプ吐出流量 $\hat{d}_1$ およびリアポンプ吐出流量 $\hat{d}_2$ は，ポンプからシリンダ内に送られる油の流量を表し，次式で計算される．

$$\text{ポンプ吐出流量} = \text{エンジン回転数} \times \text{ポンプ容量} \quad (1)$$

2) ポンプから吐出されたシリンダに供給する作動油の圧力.
3) シリンダのボトム室に流入または流出する作動油の圧力．シリンダは，シリンダチューブ，内部を移動するピストン，およびピストンに接続されたロッドから構成される．ピストンは，シリンダ内部をヘッド室とボトム室に区画する．ボトム室に作動油が流入し，ヘッド室から流出することでシリンダが伸長する．ボトム室から作動油が流出し，ヘッド室に流入することでシリンダが収縮する.
4) シリンダのヘッド室に流入または流出する作動油の圧力.

表 1　使用する操作量，状態量

ラベル	操作/状態名	単位
-	ブームレバーレート	%
-	アームレバーレート	%
-	ブームシリンダ速度	mm/s
-	アームシリンダ速度	mm/s
-	ブームシリンダ長	mm
-	アームシリンダ長	mm
-	ブームシリンダ有効力	kN
-	アームシリンダ有効力	kN
-	バケットシリンダ有効力	kN
$\hat{d}_1$	フロントポンプ吐出流量	cm^3/min
$\hat{d}_2$	リアポンプ吐出流量	cm^3/min
$\hat{d}_3$	フロントポンプ負荷出力	$MPa{\cdot}cm^3/min$
$\hat{d}_4$	リアポンプ負荷出力	$MPa{\cdot}cm^3/min$
$\hat{d}_5$	操作開始からの経過時間	s
$\hat{d}_6$	バケット刃先位置 X	mm
$\hat{d}_7$	バケット刃先位置 Z	mm
$\hat{d}_8$	X 軸方向のバケット刃先速度	m/s
$\hat{d}_9$	Z 軸方向のバケット刃先速度	m/s
$\hat{d}_{10}$	目標刃先速度	m/s
$\hat{d}_{11}$	刃先位置 Z の相対値	mm

また，フロントポンプ負荷出力 $\hat{d}_3$ およびリアポンプ負荷出力 $\hat{d}_4$ は，ポンプ吐出流量とポンプ圧を用いて次式で計算される．

$$\text{ポンプ負荷出力} = \text{ポンプ吐出流量} \times \text{ポンプ圧} \quad (2)$$

操作開始からの経過時間 $\hat{d}_5$ は，操作データの時間的変化を捉えるための車体状態である．さらに，バケット刃先の位置や速度も重要な車体状態として使用する．例えば，バケット刃先位置 X $\hat{d}_6$ およびバケット刃先位置 Z $\hat{d}_7$ は，各軸方向のバケット刃先の座標を表し，それらの情報から各軸方向の刃先速度 $\hat{d}_8$ および $\hat{d}_9$ が次式で計算できる．

$$\text{X 軸方向のバケット刃先速度} = x_t - x_{t-1} \quad (3)$$

$$\text{Z 軸方向のバケット刃先速度} = z_t - z_{t-1} \quad (4)$$

刃先位置の相対値 $\hat{d}_{11}$ は，目標刃先高さと現在の刃先位置の差分を表し，操作方向を示す重要な指標である．これらの特徴量は，空中すきとり操作における精度向上に寄与すると考えられる．

4　提案手法

RGAT では学習中に運用環境で方策を動かすことで学習用の実軌道を獲得する必要がある．しかし，本研究が対象とする油圧ショベルの自動操作タスクの特徴として，運用環境の実軌道 $\boldsymbol{\tau}_{\text{real}}$ の収集は安全面の観点から未学習や学習途中のエージェントによる操作では行えないことが挙げられる．そのため，学習途中のターゲット方策 π_θ を用いた $\boldsymbol{\tau}_{\text{real}}$ の収集ができず，RGAT をそのまま適用することができない．そのため，事前に人手で収集した操作データを $\boldsymbol{\tau}_{\text{real}}$ として使用するよう RGAT のアルゴリズムを変更する．変更したアルゴリズムをアルゴリズム 2 に示す．また変更したアルゴリズムによる，行動変換モデルおよびターゲット方策学習時の流れをそれぞれ図 4，図 5 に示す．

提案手法は，1) エキスパートによる運用環境での操作データ $\boldsymbol{\tau}_{\text{real}}$ の収集，2) 収集した操作データでのターゲット方策 π_θ の教師あり学習，3) 操作データと学習環境を用いた行動変換モデル g_ϕ の学習，4) 行動変換モデルと学習環境を用いたターゲット方策 π_θ の学習の 4 つのステップで構成されている．収集した操作データでのターゲット方策の教師あり学習は先行研究 [6] でも採用されていた事前学習方法である．

提案手法の行動変換モデル g_ϕ は学習環境の状態 s_t^{sim} と操作データ $\boldsymbol{\tau}_{\text{real}}$ の行動 a_t を受け取り，運用環境と学習環境で次状態が等しくなるような行動の補正値 δ_t を出力する．つまり行動変換モデルは，運用環境で行動 a_t を行った場合と，学習環境で行動 $a_t + \delta_t$ を行った場合の次状態が等しくなる行動の補正値 δ_t の獲得を目指す．ただし，本研究では $a_t + \delta_t$ がレバーレートの取りうる範囲を超える場合，レバーレートの範囲内に収まるよう最小値または最大値に切り詰める．この行動変換モデル g_ϕ の学習は，事前収集された操作データ $((s_t, a_t), s_{t+1})$ により計算される次式の報酬を利用した方策改善手法により行われる．

$$r_t^{\text{AT}} = -\|s_{t+1} - s_{t+1}^{\text{sim}}\|^2 - \alpha\|\delta_t\|^2 \quad (5)$$

ただし，$s_{t+1}^{\text{sim}} = f_{\text{sim}}(s_t^{\text{sim}}, a_t + \delta_t)$ はシミュレータ上での次状態を表し，f_{sim} はシミュレータを表す．また報酬の第 2 項は行動の補正値に対する正則化項であり，α は正則化係数である．正則化項の導入は，行動変換モデル g_ϕ が操作データ τ_{real} に過学習することを防ぐためである．RGAT と異なり，提案手法の行動変換モデル g_ϕ の学習には順方向遷移モデルを利用しない．本研究のタスクであるチューニング操作の初期状態はパターンが少なく，各初期状態から開始する操作データ $\boldsymbol{\tau}_{\text{real}}$ が存在する．よって収集した操作データのみで行動変換モデルの学習が可能となっている．

行動変換モデルの学習後，行動変換モデルと学習環境を用いたターゲット方策 π_θ の学習が行われる．この学習は RGAT でのターゲット方策の更新と同じく，タスクにおける報酬を利用した方策改善手法により行われる．本研究が対象とする空中すきとり操作では，報酬 R を次式のように計算する．

$$r_t = k \times V_{t+1}^{\text{X}} - |z_{t+1} - z_0| \quad (6)$$

ただし V_{t+1}^{X} は式 (3) に示す X 軸方向のバケット刃先速度，z_{t+1} は $t+1$ ステップ目での Z 座標，z_0 は初期姿勢の Z 座標である．また定数 k は 100 に設定した．この報酬関数を最大化するようにターゲット方策を学習することで，刃先を一定の高さに保ちつつ可能な限り速く引く操作を実現することを目指す．

5　シミュレータとデータセット

5.1　学習環境および運用環境

本研究で扱う学習環境には，AGX Dynamics 2.37.3.4 [7] と MATLAB&Simulink R2022a [8] を使用した．このシミュレータはシンプルなリアルタイムモデルの実行に特化している一方で，細かい物理挙動を十分に表現できないという性質がある．また実験の安全性を担保するため，運用環境は実際の油圧ショベルではなくより高精度なシミュレータである Simcenter Amesim 2021.1 [9] を使用した．このシミュレータは高度なモデリングや解析機

アルゴリズム 2 提案手法

Input: ターゲット方策 π_θ，行動変換モデル g_ϕ，方策改善手法 `optimize1`，`optimize2`, 正則化係数 α

1: 運用環境の軌道 $\boldsymbol{\tau}_{\text{real}} \leftarrow \{((s_0, a_0), s_1), ((s_1, a_1), s_2), \dots\}_{\text{real}}$ を人手で収集
2: ターゲット方策を軌道 $\boldsymbol{\tau}_{\text{real}}$ を用いた教師あり学習で事前学習
3: 式 (5) の報酬 r_t^{AT} を用いて，学習環境で `optimize1` により行動変換モデル g_ϕ を学習
4: タスクにおける報酬 r_t を用いて，学習環境で `optimize2` によりターゲット方策 π_θ を学習

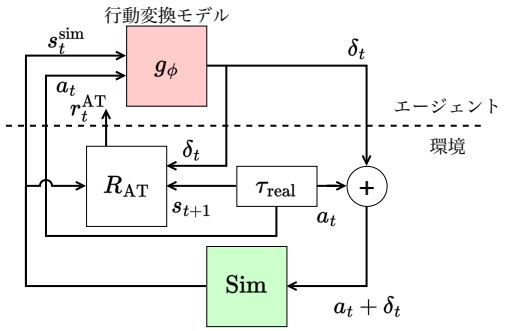

図 4　提案手法における行動変換モデルの学習

能が搭載されており, マルチフィジックスモデルを統合的に扱うことができる．一方で，シミュレーションの実行に多くの時間がかかる．

行動変換モデルの学習では，エージェントは操作開始位置で静止した状態から操作を開始する．この操作開始位置は更新時に利用する操作データの操作開始位置に合わせて設定される．操作データが終了した場合にエピソードを終了し，次の操作データの操作開始位置に刃先を移動した後に次のエピソードを開始する．ターゲット方策，すなわち自動操作モデルの学習では，行動変換モデルと同様エージェントは操作開始位置で静止した状態から操作を開始する．各エピソードは次のいずれかの条件を満たした場合に終了する．

- ステップ数が 100 より大きくなる
- 刃先位置 X が目標終了刃先位置を超える

この条件は熟練運転手が実際に行った操作データをもとに決定した．シミュレーション中の 1 ステップは，運用環境での 0.05 秒に相当する．熟練運転手の場合，5 秒間で 1 回の操作を完了できることから，運用環境の 5 秒に相当する 100 ステップを 1 エピソードの最大ステップ数とした．また，目標終了刃先位置は操作完了時の刃先位置 X の目標値であり，熟練運転手が操作を終えた座標を参考に 5400 に設定した．

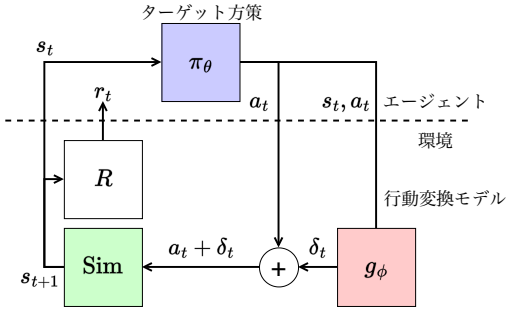

図 5　提案手法におけるターゲット方策の学習

5.2　データセット

運用環境が実環境である場合，熟練運転手が実機で操作を行うことで収集した操作データを使用することが可能である．しかし，本研究の運用環境は高精度なシミュレータの環境であるため，熟練運転手が実際に操作することは不可能である．そこで，本研究で用いるデータセットは次のように取得した．

まず，2 つのレバーのうちアームレバーは実機における熟練運転手の操縦データとする．一方で，ブームレバーについてはルールベースで設計された制御器モデルを活用しブームレバー操作量を生成している．このモデルはあらかじめ設定した基準面にバケット刃先が触れないように刃先高さをフィードバック制御するモデルである．今回の空中すきとり操作では，すきとり軌道の水平面を基準面に設定することで，アームレバーに合わせて自動でブームレバーを制御できる．この操作により得られたデータを本研究では運用環境の操作データとする．

この操作データは油圧ショベルの初期姿勢を変更しながら複数取得した．バケットの刃先位置の X 座標が $x = 9000, 9500$, Z 座標が $z = -300, 0, 300$ となるよう油圧ショベルの初期姿勢を設定した．これらの組み合わせで得られる 6 種の初期姿勢に対し，対応する基準面に沿ったすきとり操作を複数回行いデータセットを取得した．

また学習に利用する操作データとして，すきとり操作が 5 秒以内，すなわち 100 ステップ以内で刃先位置 X が目標終了刃先位置を超えるデータのみを用いた．これは，自動操作モデルのタスクである空中すきとり操作が可能な限り速く行うことを要求されているためである．

収集したデータセットは空中すきとり操作 140 回分の操作データが存在する．うち 112 回分の操作データを行動変換モデルの学習に，28 回分の操作データを行動変換モデルの性能検証に使用した．データ収集時，熟練運転手は指定された初期姿勢から操作を開始するが，初期姿勢を作るときに一定のずれが許容されるため，各エピソードの初期姿勢は完全には一致しない．そのため，自動操作モデルの強化学習を行う際のエージェントの初期状態には，各初期姿勢のデータセットの中からランダムに選択した 1 操作分の初期状態を代表値として使用した．また学習の際データセットは正規化して使用した．

6　評価実験

本節では，評価実験の実験設定と実験結果に対する評価および考察について述べる．

6.1　実験設定

モデル構造　行動変換モデルと自動操作モデルは，どちらも全結合のニューラルネットを用いた．いずれも中間層 4 層，各層 64 ユニットである．活性化関数は中間層で ReLU 関数，出力層で恒等関数を使用している．

学習設定　行動変換モデルと自動操作モデルに対する方策改善手法には，どちらも Trust Region Policy Optimization (TRPO) [10] を用いた．各モデルの学習設定を表 2 に示す．また，事前学習の設定を表 3 に示す．これらの学習設定は，表 2 の学習エピソード数を除き先行研究 [6] と同一である．

表 2　TRPO の学習設定

	行動変換モデル	自動操作モデル
学習エピソード数	10000	15000
割引率	0.99	
更新間隔	12000 ステップ	
Advantage 関数の推定方法	Generalized Advantage Estimation (λ=0.97)	
信頼半径 δ	0.01	
共役勾配法の反復	10 回	
直線探索の反復	10 回	

表 3　事前学習の学習設定

学習エポック数	2000
バッチサイズ	32
損失関数	平均二乗誤差
最適化手法	Adam
Adam のパラメータ	(β_1,β_2)=(0.9,0.999)
学習率	0.001

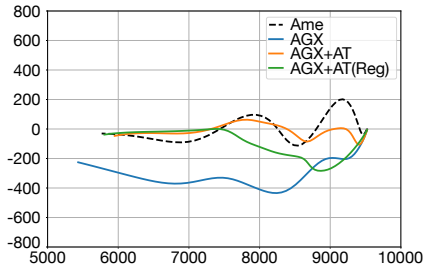

図 6　同一のレバー操作から得られる刃先軌道の比較

6.2　実験 1：行動変換モデルの評価

本実験では行動変換モデルの学習時の報酬（式 (5)）の正則化係数について $\alpha = 0$ の場合と $\alpha = 1/1000$ の 2 通りの設定で学習させた．これらの設定において学習した 2 種類の行動変換モデルを，同じレバー操作を変換した際の刃先軌道を比較することで評価する．具体的には，学習に使用していない性能検証用の 28 回分の操作データのレバー操作に対して，次の条件下の環境それぞれで空中すきとり操作を行った時の刃先軌道を比較する．

- 運用環境での刃先軌道（Ame；正解軌道）
- 学習環境での刃先軌道（AGX）
- 行動変換モデル ($\alpha = 0$) を用いて変換したレバー操作の学習環境での刃先軌道（AGX+AT）
- 行動変換モデル ($\alpha = 1/1000$) を用いて変換したレバー操作の学習環境での刃先軌道（AGX+AT (Reg)）

6.2.1　定性評価

定性評価では刃先軌道をプロットし，正しい刃先軌道との誤差を視覚的に確認する．各環境における典型的な 1 試行で得られた刃先軌道を図 6 に示す．この一例は初期姿勢が (x,z)=(9500,0) に最も近いデータである．

この図は右から左にかけてすきとり操作が進んでいる．運用環境 (Ame) は操作データに対するプロットそのものであり，このプロットが基準の刃先軌道となる．この基準からずれているほど運用環境との差異があることがわかる．学習環境での刃先軌道 (AGX) に注目すると，基準に比べて下方向にずれていることが確認できる．このずれに比べ，2 通りの設定で学習させた行動変換モデルをそれぞれ用いた場合の刃先軌道 (AGX+AT, AGX+AT (Reg)) では，どちらの場合でも基準からのずれが小さくなっていることが確認できる．一部ずれが大きくなっているステップも存在するが，すきとり操作の後半になるにつれてずれの度合いが小さくなるように補正されていることがわかる．以上の結果より，行動変換モデルを用いることで，運用環境と学習環境の差異を低減できているといえる．

表 4　検証用データセット内のレバー操作量から得られる刃先軌道と基準の刃先軌道とのユークリッド距離

	平均値	中央値
AGX	253mm	247mm
AGX+AT	243mm	134mm
AGX+AT (Reg)	225mm	169mm

6.2.2　定量評価

定量評価では刃先軌道のずれを，時間ステップごとの X, Z 座標平面上でのユークリッド距離を用いて評価する．ここでは運用環境 (Ame) の刃先軌道を基準に，他 3 つの環境の刃先軌道それぞれのユークリッド距離の平均値と中央値を計算する．これを 28 回分の操作データ全てで行った平均値と中央値の結果を表 4 に示す．この結果から，行動変換モデルを用いた学習環境の方が，用いない場合に比べて運用環境での刃先軌道とのずれが小さいことが確認できる．

6.3　実験 2：学習した自動操作モデルの評価

本実験では実験 1 で学習済みの 2 種の行動変換モデルをそれぞれ用いた学習環境上で，自動操作モデルを学習させた．比較対象として行動変換モデルを用いていない学習環境上でも自動操作モデルを学習させた．また，学習時の初期姿勢は，X 軸は 9500 に固定し，Z 軸は $z = -300, 300$ を交互に用いた．複数の初期姿勢から学習させる理由は，幅広い姿勢に対応できる自動操作モデルを獲得するためである．検証時の初期姿勢は，X 軸は 9500 に固定し，Z 軸は 21 種類の Z 座標 (-500 から+330 までをおおむね等間隔に分割した値) を用いた．このような検証設定にした理由は，学習に用いなかった初期姿勢にも自動操作モデルが対応できるかを確認するためである．本実験での行動変換モデルの評価の方法としては，次の自動操作モデルを運用環境で検証した時の刃先軌道を比較する．

- 学習環境で直接学習した自動操作モデル (RL)
- 行動変換モデル ($\alpha = 0$) を用いて学習した自動操作モデル (RL+AT)
- 行動変換モデル ($\alpha = 1/1000$) を用いて学習した自動操作モデル (RL+AT (Reg))

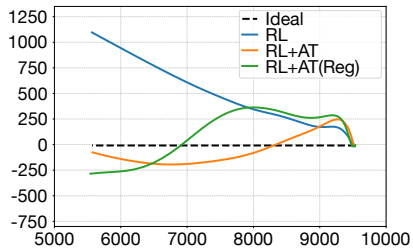

図 7　運用環境での典型的な刃先軌道の比較

表 5　運用環境における各自動操作モデルの刃先軌道と理想刃先軌道のユークリッド距離

	最大値	平均値
RL	1150mm	371mm
RL+AT	266mm	110mm
RL+AT (Reg)	386mm	198mm

これらの刃先軌道に加え，比較のために初期位置から水平に引かれた理想刃先軌道 (Ideal) を用いる．

6.3.1　定性評価

定性評価では刃先軌道をプロットし，理想刃先軌道との誤差を確認する．それぞれの環境で学習させた自動操作モデルを運用環境で検証した際に得られた刃先軌道の例を図 7 に示す．この一例は初期姿勢が (x, z)=(9500, 0) に最も近いデータである．

行動変換モデルを用いず，学習環境で学習した自動操作モデル (RL) は，運用環境で検証した際に理想刃先軌道から大きくずれる．一方で，2 通りの設定で学習させた行動変換モデルをそれぞれ用いて学習した自動操作モデル (RL+AT, RL+AT (Reg)) は，運用環境で検証した際に，行動変換モデルを用いずに学習した場合と比べて理想刃先軌道からのずれが小さくなっていることが確認できる．以上の結果から，行動変換モデルにより運用環境と学習環境の差異を低減できていることが確認できた．

6.3.2　定量評価

定量評価では刃先軌道のずれを同一の X 座標，Z 座標のユークリッド距離を用いて評価する．ここでは理想刃先軌道 (Ideal) を基準に，他 3 つの刃先軌道それぞれのユークリッド距離の最大値とエピソード平均値をそれぞれ計算する．21 種類の初期姿勢に対して得られた最大値および平均値を表 5 に示す．

この結果から，行動変換モデルを用いて学習環境で学習した自動操作モデルの方が，用いない場合に比べて運用環境での刃先軌道とのずれを小さくすることが確認できる．また，$\alpha = 0$ の場合の報酬関数で学習した行動変換モデルの方が $\alpha = 1/1000$ の場合の報酬関数で学習した行動変換モデルよりも良い結果になっている．この理由として，式 (5) の報酬に含まれる正則化項が悪影響を与えた可能性や，α の調整が必要だった可能性が考えられる．また，$\alpha = 0$ の設定で学習した行動変換モデルを用いた場合，理想刃先軌道とのユークリッド距離の最大値が 266mm であり，操作の速さを優先する燃費測定試験に要求される基準の 300mm 以内 [11] を実現した．

7　結論

本研究では油圧ショベルの空中すきとり操作を行うための自動操作モデルを，運用環境と異なる学習環境で学習するための Sim-to-Real Transfer の方法を提案した．提案手法は運用環境と学習環境の次状態が近づくように行動を変換する行動変換モデルを導入した Sim-to-Real 強化学習手法である RGAT に基づいている．提案手法ではこの RGAT に対して，学習中の自動操作モデルを実環境で動かす必要のある軌道収集のプロセスを削除し，事前収集したエキスパートの軌道のみを利用するよう変更した．評価実験では提案手法により得られた行動変換モデルが環境の差異を低減することを確認し，自動操作モデルが運用環境においても要求基準を満たす性能で空中すきとり操作を実現することを確認した．

今後の展望としては，運用環境を高精度シミュレータである Simcenter Amesim の環境から，実環境に変更し，自動操作モデルを実運用化することが挙げられる．また，油圧ショベルや空中すきとり操作だけでなく，様々な建設機械やタスクに対応した行動変換モデルの開発が求められる．

参考文献

[1] 国土交通省. 最近の建設業を巡る状況について, 2023, https://www.mlit.go.jp/policy/shingikai/content/001633500.pdf.

[2] Wenshuai Zhao, Jorge Pena Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 737–744, 2020.

[3] Haresh Karnan, Siddharth Desai, Josiah P. Hanna, Garrett Warnell, and Peter Stone. Reinforced grounded action transformation for sim-to-real transfer. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4397–4402, 2020.

[4] Keita Matsumoto, Atsushi Yamaguchi, Takahiro Oka, Masahiro Yasumoto, Satoru Hara, Michitaka Iida, and Marek Teichmann. Simulation-based reinforcement learning approach towards construction machine automation. In Furuya Hiroshi, Tateyama Kazuyoshi, and Osumi Hisashi, editors, *Proceedings of the 37th International Symposium on Automation and Robotics in Construction (ISARC)*, pp. 457–464, Kitakyushu, Japan, October 2020. International Association for Automation and Robotics in Construction (IAARC).

[5] 泉翔太, 谷島諒丞, 全邦釘. 強化学習を用いたバックホウの掘削動作生成. AI・データサイエンス論文集, Vol. 1, No. J1, pp. 307–312, 2020.

[6] 榊原龍志. 強化学習を用いた油圧ショベルの操作モデルの学習. Master's thesis, 横浜国立大学 大学院環境情報学府, 2023. 修士論文.

[7] Algoryx Simulation AB. AGX dynamics documentation, 2023, https://www.algoryx.se/documentation/complete/agx/tags/latest/doc/UserManual/source/index.html.

[8] MathWorks. Documentation - MATLAB & Simulink, 2023, https://jp.mathworks.com/help/.

[9] Siemens. Simcenter Amesim, 2025, https://plm.sw.siemens.com/ja-JP/simcenter/systems-simulation/amesim/.

[10] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, July 2015. PMLR.

[11] 一般社団法人日本建設機械施工協会. 土工機械－エネルギー消費量試験方法－油圧ショベル JCMAS H 020:2014, 2014, https://jcmanet.or.jp/jcm/wp-content/uploads/2024/01/JCMAS_H_020.pdf.