# モバイルマニピュレータのための タスクに関連する視点・領域に着目した頑健な模倣学習

○石田 裕太郎　野口 裕貴　金井 嵩幸　新谷 和宏　尾藤 浩司（トヨタ自動車株式会社）

本研究では，モバイルマニピュレータにおける模倣学習による行動獲得において問題となる遮蔽やドメインシフトに対し，タスクに関連する視点・領域に着目した頑健な手法を提案する．提案手法は複数視点とそれら領域に対する attention 機構と幾何学模様による高速かつ低計算コストな水増しにより構成される．これにより，先行研究と比べ，遮蔽やドメインシフトが発生する環境下で 13.4 ポイント以上のタスク成功率の向上を確認した．

## 1.　はじめに

近年，様々な場所で多様なタスクを実行できるモバイルマニピュレータが注目を浴びている [1]．それに関連した Robot Learning，認識 [2]，制御 [3]，シミュレーション [4]，ロボット競技会 [5] など，社会実装を模索する広範な研究開発が行われている．

Robot Learning では，視覚観測に基づいて次の行動を予測する visuomotor と呼ばれる手法が一般的に用いられる．しかし，モバイルマニピュレータにこの手法を用いる場合，2 つの問題が存在する．1 つ目は，モバイルマニピュレータに搭載されるセンサを用いる場合，視覚観測において遮蔽が起こりやすいことである．例えば，図 1 左に示す pick タスクでは，ヘッド視点は自身により遮蔽される．一方で，図 1 右に示す place タスクでは，ハンド視点は把持した物体により遮蔽される．したがって，複数視点を用い，それらの中から状況に応じてタスクに関連する視点に着目し（遮蔽のない視点に着目し），遮蔽に対応できなければならない．2 つ目は，移動台車を有さないマニピュレータに比べ，モバイルマニピュレータは様々な場所で働くため，視覚観測においてドメインシフトが起こりやすいことである．例えば，図 2 左に示す学習データを収集した In Distribution (ID) 環境に対し，図 2 中に示すタスクに関係ない物体が存在する，図 2 右に示す背景（家具・壁紙）が異なる，これら Out Of Distribution（OOD）環境があり，ドメインシフトが起こる．したがって，視覚観測の中でタスクに関連する領域に着目し，タスクに関連しない領域には過剰に反応しないようにし，ドメインシフトに対応できなければならない．

先行研究 [6] は，ドメインシフト（ロボットの位置，タスクに関係ない物体，背景の変化）に対して，ハンド視点が第三者視点より一意に頑健であると仮定し，着目すべきハンド視点の情報は全て使い，第三者視点の情報は画像エンコーダの後段に Variational Information Bottleneck（VIB）を追加し最小限にフィルタリングした．しかしながら，前述した 1 つ目の問題の通り，タスクに関連する視点は状況に応じて変わるため，この一意に仮定する手法は遮蔽には十分に対応できない．また，前述した 2 つ目の問題の通り，一意に仮定された視点に着目するだけではドメインシフトには対応できない．

他の先行研究 [7, 8] は，長時間の操縦によるデモンストレーションの収集と，拡散モデルを用いた水増しにより，大規模なデータセットを収集した．これを spatial attention を含む方策で学習することで，タスクの関連する領域に着目することができた．しかしながら，この研究では遮蔽に対応することは検討されていない．

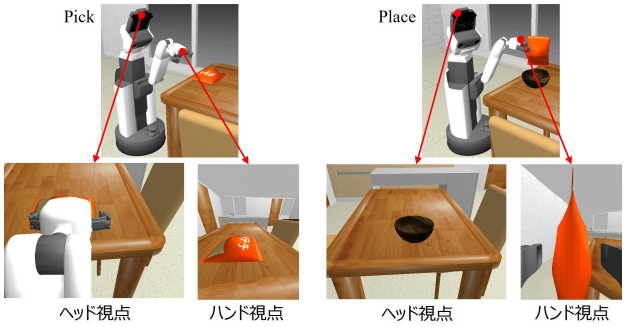


図 1　視覚観測における遮蔽


図 2　視覚観測におけるドメインシフト

本研究では，複数視点を有すモバイルマニピュレータにおいて，タスクに関連する視点・領域に着目したロバストな模倣学習手法を提案する．まず，複数視点にまたがるタスクに関連する領域に着目するため attention 機構を提案する．次に，attention 機構の学習を容易にするため，幾何学模様による高速かつ低計算コストなタスクに関連しない領域の水増しを提案する．これにより，遮蔽やドメインシフトに対する頑健性が向上する．提案手法と先行研究を比較した結果，遮蔽やドメインシフトが発生する環境下で 13.4 ポイント以上のタスク成功率向上を確認した．

## 2.　関連研究

### 2.1　モバイルマニピュレータ

モバイルマニピュレータは，移動台車を使って様々な場所に移動してタスクを行える．また，移動台車を使って作業中に姿勢を変えることができ，多様なタスクを実行できる可能性がある [1, 3]．その知能は，認知，判断，操作に大別され，人間の手作業で設計・実装されたアルゴリズムで実現されている [2]．しかし，熟練

したエンジニアによる設計・実装には長い時間を要する [5]．また何かを調理するようなタスクに対するアルゴリズムを設計・実装することは，長い時間をかけても実現が困難なことがある [9]．そのため，前述とは別のアプローチとして，Robot Learning のような模倣や試行錯誤による行動の獲得が期待されている．

### 2.2 Robot Learning

近年は模倣学習による行動の獲得に関する研究が活性化している [6, 7, 8, 9]．しかし，1. 章で述べたように，筆者の知る限り遮蔽やドメインシフトの影響を受けやすいモバイルマニピュレータのための頑健な手法の研究はない．

近年の画像・言語分野における基盤モデルの成功を受け，Robot Learning もアーキテクチャおよびデータ量の観点で模索している．アーキテクチャの面では，Transformer[7, 9] や，条件付き拡散モデル [10, 11] に基づいた研究がある．データ量の面では，長時間の操縦によるデモンストレーションの収集が試みられている [7]．しかし，操縦で全ての状況におけるデモンストレーションを収集することは現実的ではない．そのため，例えば基盤モデルでも用いられる spatial attention[12] を用いることで，タスクに関連しない領域のドメインシフトの影響を軽減し，データ効率を向上する必要がある．

## 3. 定義

### 3.1 タスクに関連する領域と関連しない領域

ここではモバイルマニピュレータが環境と相互作用することによりタスクを解決することを考える．ここで，“タスクに関連する領域” とは，視覚観測内のタスクに関連する物体の領域と定義する．“タスクに関連しない領域” とは，その他の領域と定義する．例えば，図 2 の赤い缶を pick するタスクでは，タスクに関連する領域とは赤い缶に対応する画素であり，タスクに関連しない領域とはその他の画素である．

### 3.2 模倣学習

ここではモバイルマニピュレータがデモンストレーション $\mathcal{D} = \{(o_{\text{expert}}{}_t^{(n)}, a_{\text{expert}}{}_t^{(n)})_{t=0}^{T^{(n)}}\}_{n=0}^{N}$ を模倣し，機械学習により方策 $\pi(a|o)$ を獲得する模倣学習を考える．$o_{\text{expert}}$ と $a_{\text{expert}}$ はモバイルマニピュレータの観測と行動であり，$T$ は各エピソード長であり，$N$ はエピソード数である．方策はデモンストレーションを用いた教師あり学習により $o_{\text{expert}}$ が入力された時に予測した行動 $a_{\text{pred}}$ と $a_{\text{expert}}$ の誤差を最小化することにより得る．テスト時には，方策に現在の観測 $o$ を入力し次の行動 $a$ を予測することを繰り返すことで動作する．

### 3.3 観測と行動

ここでは図 1 に示す複数視点を有するモバイルマニピュレータを考える．観測は，ハンド視点 $o_h$，ヘッド視点 $o_f$，手先姿勢 $o_p$ である．行動 $a$ は手先の相対移動量である．低レベルコントローラは方策から $a$ を受け取り，それを関節位置指令値に変換するために逆運動学を解き，モバイルマニピュレータが動作する．

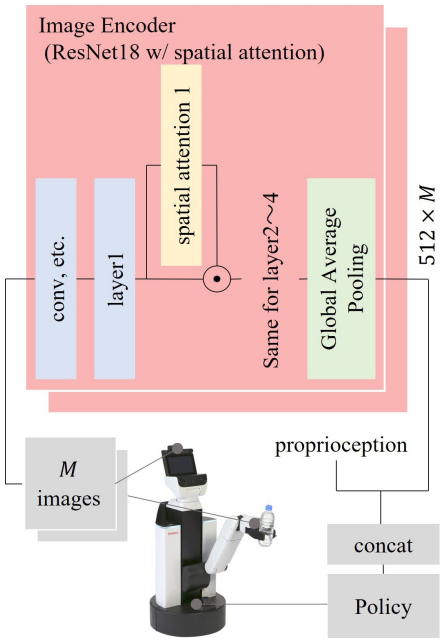


図 3　Attention 機構を組み込んだアーキテクチャ

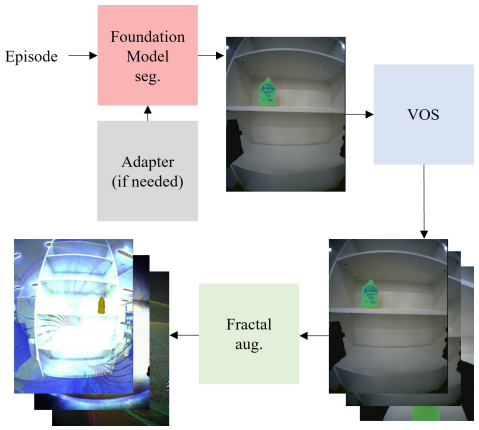


図 4　高速かつ低計算資源な水増し

## 4. 提案手法

### 4.1 複数視点とその領域に対する attention 機構

提案手法は，状況に応じてタスクに関連する視点に着目し遮蔽に対する頑健性を向上するため，図 3 に示すように，共有画像エンコーダの代わりに，$M$ 個の視点それぞれに対応するパラメタを持つ，$M$ 個の画像エンコーダを組み込む．また，提案手法は，タスクに関連する領域に着目しドメインシフトに対する頑健性を向上するため，先行研究 [7] に倣い，図 3 に示すように，画像エンコーダの layer 毎に spatial attention を組み込む．spatial attention は画像から得た特徴 $f(b, h, w, c)$ に対して，重み $w(b, h, w, 1)$ を予測し，それをチャンネル次元に展開し，$f(b, h, w, c)$ と $w(b, h, w, c)$ の内積をとる．これにより特徴量はタスクに関連する領域に重みづけされる．

提案手法はこの attention 機構を用いて，複数視点にまたがるタスクに関連する領域の情報のみを保持した

埋め込みを方策に与える．これは，1. 章で述べた先行研究とは異なり，様々な状況で発生する遮蔽とドメインシフト両者に対して頑健な手法である．

### 4.2　幾何学模様による高速かつ低計算コストな水増し

Attention 機構の学習を容易にするために，幾何学模様でタスクに関連しない領域を水増しする方法を提案する．水増しをすることで，タスクに関連しない領域は幾何学模様による大きな変化を受け，attention 機構は変化を受けていないタスクに関連する領域により着目する．先行研究 [8] とは異なり，提案手法は高速かつ低計算コストであるため，単一視点と比較してより大量の画像を扱わなければならない複数視点の水増しに適している．

提案手法は，図 4 左上に示すように，データセット $\mathcal{D}$ における $t=0$ の $o_h$ と $o_f$ のタスクに関連する領域を検出する．検出には FastSAM[13] のような基盤モデルを用い，テキストプロンプトを用いてゼロショットでオブジェクトを検出する．次に，図 4 右に示すように，タスクに関連する領域を $t=0$ から $t=T(n)$ まで追跡する．追跡には XMem[14] のような Video Object Segmentation（VOS）を用いる．これらのステップにより，全てのエピソード中の全ての時刻について，タスクに関連した領域のマスクが生成できる．最後に，図 4 左下のように，様々な環境をシミュレートするために，PixMix[15] を用いてタスクに関連しない領域に水増しを適用する．幾何学模様は画像分類器の事前学習 [16] に有用であることが証明されており，ドメインシフトへの頑健性を向上させるために，学習手順に組み込まれる．必要であれば FastSAM のテキストプロンプトは，CLIP-Adapter[17] を使ってロボットドメインのデータに適応させマスク精度を向上させることができる．

## 5.　実験・考察

### 5.1　タスクと環境

**place-banana-on-plate タスク:** 図 5 にタスクのシーケンスを時系列に示す．モバイルマニピュレータが $o_f$ で床に置かれたプレートを観測し（図 5 左），プレートに手を伸ばし（図 5 中央），バナナをプレートに置く（図 5 右）．モバイルマニピュレータがバナナを床に触れることなくプレートに置くことができればタスクの完了とした．図 7 に示すとおり，タイムステップの初期は，$o_f$ は遮蔽がなく，$o_h$ はバナナと背景を観察しているだけで遮蔽されていた．タイムステップの後期は，$o_f$ は自身により遮蔽が増加し，$o_h$ はプレートを観察でき遮蔽が軽減した．

図 7 にタスクの 2 つの環境の種類を示す．図 7 左の ID 環境は，デモンストレーションを収集したときと同じ環境である．図 7 右の OOD-floor 環境は，未知の床の模様を有す環境である．デモは ID 環境で 50 回収集し，モバイルマニピュレータ，プレートの位置を変えながら各環境で 15 回テストした．

### 5.2　実装方法

比較のために，(1) 先行研究：CNN 版 Diffusion Policy（DP）[11]，(2) 先行研究：画像エンコーダに VIB を付け水増し（Aug）を用いた CNN 版 DP[6]，(3) 提案手法：attention 機構（AM）と水増しを用いた CNN 版 DP, 3 つの手法を実装した．DP のハイパーパラメータは CNN 版の Real 環境と揃えた [11]．DP の画像エンコーダの初期重みは ImageNet で pre-training したものとした．VIB と spatila attention のアーキテクチャをそれぞれ表 1 と表 2 に示す．VIB の損失関数に用いた $\beta$ は 1.0 とした．モバイルマニピュレータとしてトヨタ Human Support Robot（HSR）[1] を用いた．

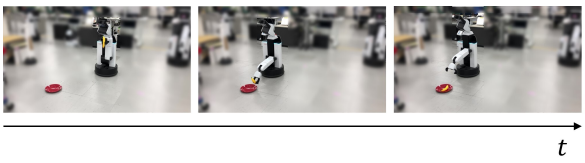


図 5　place-banana-on-plate タスクのシーケンス

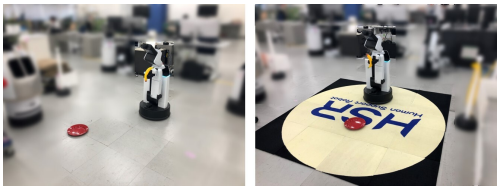

図 6　place-banana-on-plate の環境

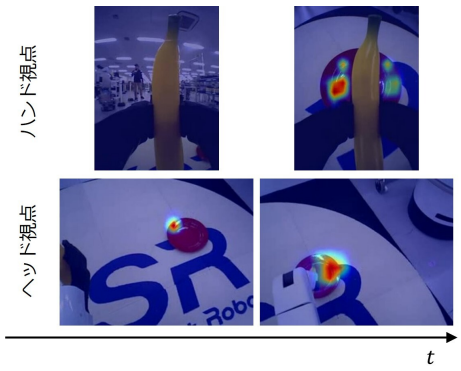


図 7　可視化した方策が着目した視点・領域

### 5.3　結果と考察

place-banana-on-plate タスクにおける提案手法と先行研究の比較を表 3 に示す．いずれの手法も 3 つの観測 $o_h$，$o_f$，$o_p$ を持つ．先行研究に従い，$o_f$ の特徴量 $z_f$ に対して VIB を適用した．

先行研究の DP と比較して，提案手法は OOD-floor 環境で 13.4 ポイント，タスク成功率を改善した．この改善は，4. 章で説明したように，attention 機構と水増しが遮蔽とタスクに関連しない領域のドメインシフトに頑健であると考えられる．

先行研究の VIB と比較して，提案手法は ID 環境で 25.6 ポイント，OOD-shelf 環境で 66.7 ポイント，タスク成功率を改善した．これは，図 7 に GradCAM[18] を用いて可視化した通り，提案手法が時々刻々と変わる遮蔽状況のデータセットからタスクに関連する視点・領域を学習でき着目できたことによる．

表 1　VIB のアーキテクチャ

| | Layer Description | Output Tensor Dim. |
|---|---|---|
| #0 | Input feature | 512 |
| | Encoder | |
| #1 | Linear + ReLU | 256 |
| #2 | Linear + ReLU | 256 |
| | $\mu$ layer | |
| #3-1 | Linear | 128 |
| | $\sigma$ layer | |
| #3-2 | Linear + Softplus | 128 |
| | Latent layer | |
| #4 | $\mu + \sqrt{\sigma} \times randn$ | 128 |
| | Decoder | |
| #5 | Linear | 512 |

表 2　spatial attention のアーキテクチャ

| | Layer Description | Output Tensor Dim. |
|---|---|---|
| #0 | Input feature | $c \times h \times w$ |
| #1 | Conv2d + ReLU | $c \times h \times w$ |
| #2 | Conv2d + Sigmoid | $1 \times h \times w$ |
| #3 | expand (Channel Dim.) | $c \times h \times w$ |

表 3　place-banana-on-plate タスクの成功率

| method | Success Rate [%] ID | OOD (floor) |
|---|---|---|
| Original [11] | **100.0** | 73.3 |
| VIB($z_f$) + Aug [6] | 66.7 | 20.0 |
| **Proposal: AM + Aug** | 93.3 | **86.7** |

## 6.　まとめ

本研究では，複数視点を有すモバイルマニピュレータにおいて，タスクに関連する視点・領域に着目したロバストな模倣学習手法を提案した．提案手法は attention 機構を組み込んでおり，水増しされたデータセットを用いて学習され，遮蔽やドメインシフトに対して最適な視点と頑健な視覚埋め込みをもたらした．

今後の課題として実験に用いるタスクや環境を増やし，より遮蔽やドメインシフトが激しい状況での提案手法の有効性を示すことが挙げられる．

### 参　考　文　献


[1] T. Yamamoto, et al.: "Development of the Research Platform of a Domestic Mobile Manipulator Utilized for International Competition and Field Test," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 7675–7682, 2018.

[2] T. Ono, et al.: "Solution of World Robot Challenge 2020 Partner Robot Challenge (Real Space)," Advanced Robotics, vol. 36, no. 17–18, pp. 870–889, 2022.

[3] T. Yamamoto, et al.: "Development of Human Support Robot as the research platform of a domestic mobile manipulator," ROBOMECH Journal, vol. 6, no. 4, 2019.

[4] Y. Matsusaka, et al.: "A Continuous Integration Based Simulation Environment for Home Support Robot and its Application to RoboCup Competition," IEEE/SICE International Symposium on System Integration, pp. 1–6, 2023.

[5] L. Contreras, et al.: "Towards general purpose service robots: World Robot Summit – Partner Robot Challenge," Advanced Robotics, vol. 36, no. 17-18, pp. 812–824, 2022.

[6] K. Hsu, et al.: "Vision-Based Manipulators Need to Also See from Their Hands," International Conference on Learning Representations, 2022.

[7] A. Brohan, et al.: "RT-1: Robotics Transformer for Real-World Control at Scale," arXiv preprint arXiv:2212.06817, 2022.

[8] T. Yu, et al.: "Scaling Robot Learning with Semantically Imagined Experience," arXiv preprint arXiv:2302.11550, 2023.

[9] Z. Fu, et al.: "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," arXiv preprint arXiv:2401.02117, 2024.

[10] M. Janner, et al.: "Planning with Diffusion for Flexible Behavior Synthesis," International Conference on Machine Learning, 2022.

[11] C. Chi, et al.: "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," Robotics: Science and Systems, 2023.

[12] M. S. Ryoo, et al.: "TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?," arXiv preprint arXiv:2106.11297, 2021.

[13] X. Zhao, et al.: "Fast Segment Anything," arXiv preprint arXiv:2306.12156, 2023.

[14] H. K. Cheng and A. Schwing: "XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model," European Conference on Computer Vision, 2022.

[15] D. Hendrycks, et al.: "PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures," Computer Vision and Pattern Recognition, 2022.

[16] H. Kataoka, et al.: "Pre-training without Natural Images," International Journal of Computer Vision, 2022.

[17] P. Gao, et al.: "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," arXiv preprint arXiv:2110.04544, 2021.

[18] R. R. Selvaraju, et al.: "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," IEEE International Conference on Computer Vision, 2017.