# マルチモーダル基盤モデルと劣モジュラ最適化に基づく移動ロボットの環境探索

Submodular Observation Poses Optimization for Mobile Robots

松尾 榛夏 Haruka Matsuo　神原 元就 Motonari Kambara　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

In this paper, we address the task where domestic service robots perform object manipulation within a domestic environment following user instructions. Since objects are often moved in daily life, it is important to efficiently perform periodic searches to obtain the latest object positions. The more images collected, the more objects can be located, but this is time-consuming. Focusing on a small number of images can reduce the time required, however, the amount of information obtained is limited by the possibility of obtaining images of walls, corridors with few objects, and so on. Most existing methods do not properly account for the possibility that everyday objects may move to different positions at different points in time. In this paper, we propose an observation pose optimization method that takes advantage of object presence maps and submodularity. The proposed method outperformed the baseline method in terms of the ratio of everyday objects that can be observed in the observations.

## 1. はじめに

ロボット技術は様々な分野で注目を集めるようになっており，特に生活やセキュリティの分野において有望な解決策として期待されている．例えば家庭用の生活支援ロボットは，高齢者や身体的支援が必要な人々の生活を支援し，日常のタスクを効果的に補助することが期待される．こうしたロボットにとって，事前に家庭環境全体に散在する物体位置を把握できれば効率的にタスクを実行できる．このようなタスクの例として，[Kaneda 24] で定義されたタスクが挙げられる．一方で，日常的な物体は日常生活において頻繁に移動され位置が変化するため，定期的な探索を効率的に行い，最新の物体位置を把握することが重要となる．

本研究の対象タスクにおいて，ユーザによって観測数が与えられた場合に，モデルはロボットができるだけ多くの日常的な物体を観測できるような，2D 平面上における有益なロボット姿勢集合を選択することが望ましい．本研究では，効果的な環境観測のためにロボットの 2D 姿勢集合の最適化を目的とした観測姿勢集合組み合わせ最適化（COPO）に焦点を当てる．COPO は，与えられた環境を観測するための最良の姿勢集合を選択することを目指したアプローチを必要とする．観測姿勢の数が増加するにつれて，観測姿勢の可能な組み合わせの数は指数関数的に増加する．一方で，[Nemhauser 78] において示されているように，劣モジュラ性を用いることで貪欲法により最適スコアの $(1-1/e)$ 近似が保証される．これは，貪欲法を用いることで最適スコアの 63%が保証され，組み合わせ爆発の問題を回避できることを意味する．

そこで，本研究では Submodular Observation Poses Optimization (SOPO) を提案する．これは家庭環境内における日常物体の観測を目的とした，劣モジュラ性を利用した観測姿勢集合の最適化手法である．本提案手法では，NP 困難である COPO を扱うため，劣モジュラ性を利用する．加えて，Open–vocabulary でマルチモーダルな 3D 特徴量およびテキストプロンプトから生成した positive object occurrence map および negative occurrence map を用いる．

連絡先: 松尾榛夏，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，haruka.matsuo-25@keio.jp

図 1: COPO の代表例.

本研究の独自性は以下である．

- 家庭環境で日常物体を観測することを目的とした，観測姿勢集合に対する劣モジュラ最適化手法を提案する．
- positive object occurrence map および negative occurrence map を扱うための Occurrence Mapping Module を導入する．これにより，日常物体や遮蔽物が存在する領域を考慮できる．

## 2. 問題設定

本研究では，効果的な環境観測を目的とした 2D 平面上でのロボット姿勢の最適化を対象とする．本タスクでは，環境内の観測可能な物体数を最大化するために，2D 平面上でのロボット姿勢を最適化することが望ましい．多くの物体を観測することは，無限の観測点を使用することにより可能になるが，この簡単な解決策は避ける必要がある．したがって，観測点の数はユーザによって指定され，固定されていると仮定する．図 1 に本タスクの代表例を示す．中央の画像は環境の 3D モデルおよび最適化された 2D 平面上のロボット姿勢を示す．中央の画像を囲む 4 枚の画像は各観測姿勢で収集された観測を示す．モデルはロボットが多くの日常物体を観測できるような有益なロボット姿勢の集合を選択する必要がある．これは典型的な組み合わせ最適化問題であり，NP 困難である．

入力は事前の探索で得られた 2D map および環境中の家具に関する点群である．出力は環境を観測するための，2D 平面上におけるロボット姿勢の集合である．ここで，本論文で使用する用語として，日常物体について，日常的に使用され生活支援ロボットによる把持及び移動が可能な物体として定義する．

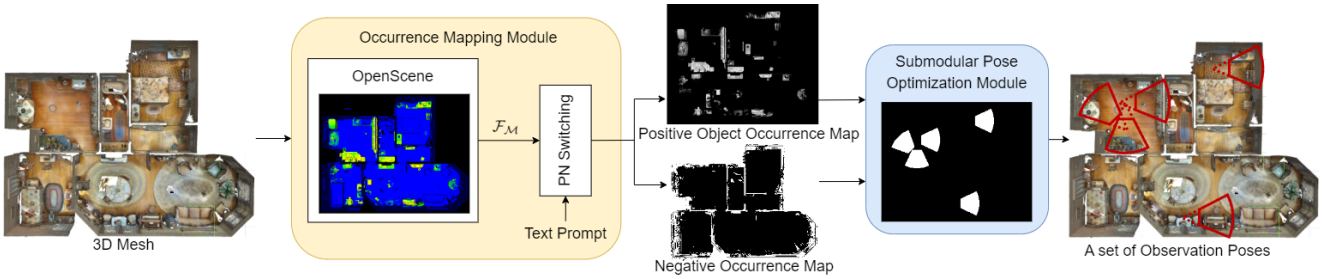


図 2: 提案手法のネットワーク構造.

具体例として，本，ペットボトル，及びマグカップが挙げられる．また，観測姿勢について，ロボットが環境を観測する位置および向きと定義する．本研究では，環境の観測に単一のカメラのみを使用し，カメラパラメータおよび環境が既知であるとする．さらに，特定の個々の物体の探索ではなく，可能な限り多くの異なる物体の観測に焦点を当てている．また，観測地点間のナビゲーションに関する軌道生成は，巡回セールスマン問題の解法のような既知の最適化手法に基づくと仮定する．

## 3. 提案手法

図 2 に提案手法のモデル構造を示す．本モデルの主要モジュールは Occurrence Mapping Module (OMM) および Submodular Pose Optimization Module (SPOM) の 2 つである．本手法は，[Sugiura 18] のような active sensing タスクと関連が深い．

モデルへの入力を 2 つ定義する．1 つ目は環境内の家具に関する点群 $\boldsymbol{x}_{\mathrm{pcl}} \in \mathbb{R}^{M\times 3}$ であり，$M$ は点群の数を表す．2 つ目は事前に探索された 2D 占有格子地図 $\boldsymbol{x}_{\mathrm{map}} \in \mathbb{R}^{r\times c}$ であり，$r$ および $c$ はそれぞれ 2D 占有格子地図の行と列を表す．

### 3.1 Occurrence Mapping Module

OMM では，日常物体が含まれている可能性が高い領域を特定するための positive object occurrence map および遮蔽物の領域を特定する negative occurrence map を生成する．COPO は，日常物体の観測数ができるだけ多くなるようなロボット姿勢を決定する組み合わせ最適化問題である．そこで，本研究では，CLIP [Radford 21] ベースのモデルを用いて共起確率を予測する．これによって，大規模なデータセットによって得られた常識的な知識を用いることが可能であるため，既存手法では扱われていなかった，非常に広い種類の物体の共起確率を扱うことが可能となる．つまり，本モジュールは COPO のための positive object occurrence map および negative occurrence map を生成する．これらのマップは Open–vocabulary なマルチモーダル 3D 点群およびテキストプロンプトから生成する．具体的には，3D 特徴量を取得するために OpenScene を使用する．

まず，$\boldsymbol{x}_{\mathrm{pcl}}$ から生成された 3D メッシュモデル内で RGB 画像を収集する．そして，事前訓練済みのセグメンテーションモデルである OpenSeg [Ghiasi 22] を用いて，RGB 画像からピクセルごとの埋め込みを計算する．次に，中間特徴量 $\boldsymbol{\mathcal{F}}^{\mathrm{2D}} = \{\boldsymbol{f}_1^{\mathrm{2D}}, ..., \boldsymbol{f}_M^{\mathrm{2D}}\} \in \mathbb{R}^{M\times C}$ を得る．ここに，$C$ は 3D 表面点の特徴次元を表し，$\boldsymbol{f}_i^{2D}$ は平均プーリングを使用して埋め込みから計算された $\boldsymbol{x}_{\mathrm{pcl}}$ のインデックス $i$ の点に対する特徴ベクトルである．その後，MinkowskiNet18A [Choy 19] を使用することにより $\boldsymbol{x}_{\mathrm{pcl}}$ から点ごとの埋め込み $\boldsymbol{\mathcal{F}}^{\mathrm{3D}} = \{\boldsymbol{f}_1^{\mathrm{3D}}, ..., \boldsymbol{f}_M^{\mathrm{3D}}\} \in \mathbb{R}^{M\times C}$ を得る．さらに，事前に定義された $N$ 個の物体クラスラベルの集合に対して，CLIP テキストエンコーダを使用して，埋め込み $\boldsymbol{T} = \{\boldsymbol{t}_1, ..., \boldsymbol{t}_N\} \in \mathbb{R}^{N\times C}$ を計算する．各 3D 点に対して，コサイン類似度を使用してアンサンブルスコア $\boldsymbol{s}^{\mathrm{2D}} = \max_n(\cos(\boldsymbol{f}^{\mathrm{2D}}, \boldsymbol{t}_n))$, $\boldsymbol{s}^{\mathrm{3D}} = \max_n(\cos(\boldsymbol{f}^{\mathrm{3D}}, \boldsymbol{t}_n))$ をそれぞれ取得する．そして，$\boldsymbol{s}^{\mathrm{2D}}$ および $\boldsymbol{s}^{\mathrm{3D}}$ のうち最も高い値を持つ特徴量である点群特徴量 $\boldsymbol{\mathcal{F}}^{\mathcal{M}}$ を取得する．

次に NP Switching を行う．この処理では，まず与えられたテキストプロンプトに対応する場所が高いアンサンブルスコアを持つ 3D メッシュを $\boldsymbol{\mathcal{F}}^{\mathcal{M}}$ およびプロンプトから取得する．そして，3D メッシュを 2D 画像に変換し，positive object occurrence map $\boldsymbol{u}_{\mathrm{obj}} \in \mathbb{R}^{r\times c}$ を取得する．さらに，3D メッシュの値の無い領域を遮蔽物として考慮し，2D 画像に基づいて negative occurrence map $\boldsymbol{u}_{\mathrm{neg}} \in \{0,1\}^{r\times c}$ を取得する．

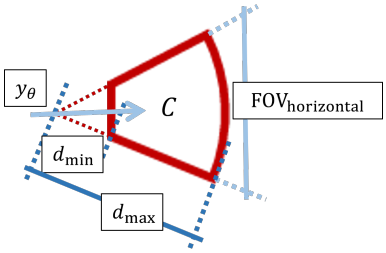


図 3: カメラのモデル.

### 3.2 Submodular Pose Optimization Module

SPOM はカバレッジを最大化する観測姿勢集合を生成する．本モジュールの入力は $\boldsymbol{u}_{\mathrm{neg}}$ および $\boldsymbol{x}_{\mathrm{map}}$ であり，出力は観測姿勢集合 $A_K$ である．ここに，$K$ は集合のサイズを表す．

まず，本研究で使用される図 3 に示すカメラのモデルを説明する．$d_{\mathrm{min}}$ および $d_{\mathrm{max}}$ は各観測姿勢のカバレッジ領域の制限を表す．それぞれ，カメラが物体を観測可能な最小および最大距離である．また，$\mathrm{FOV}_{\mathrm{horizontal}}$, $y_\theta$, $C$ はそれぞれカメラの水平 FOV，観測姿勢の方向，およびある観測地点がカバーする観測領域を表す．

観測姿勢 $y$ が導入されたとき，そのカバレッジ $y_{\mathrm{cov}}$ は次のようにモデル化される：

$$y_{\mathrm{cov}} = C(y) = \begin{cases} 1 & \text{if coverage area} \\ 0 & \text{otherwise} \end{cases}$$

観測姿勢集合 $A$ から得られるカバレッジを $C(A)$ と定義する．ここで，観測姿勢が入力 $\boldsymbol{x}_{\mathrm{map}}$ の観測姿勢候補集合 $V$ の 1 つであると仮定する．$A$ および $f_{\mathrm{cov}}(A)$ をそれぞれ既に選択された姿勢の集合および $A$ から得られるカバレッジとする．$f_{\mathrm{cov}}(A)$ は以下のように定義する：

$$f_{\mathrm{cov}}(A) = \sum_{p\in\bigcup_{j\in A} P_j} c_p$$

ここに，$j$ および $p$ はそれぞれ $A$ 内の観測姿勢および $P_j$ 内の各姿勢を表し，$P_j$ および $c_p$ はそれぞれ $j$ の集合および姿勢 $p$ の重みを表す．

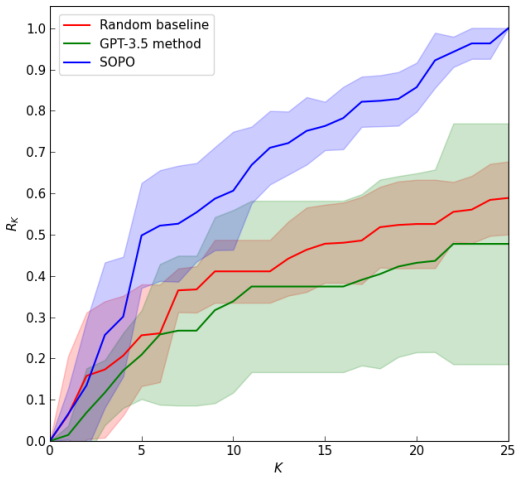


図 4: ベースライン手法群および提案手法の比較.

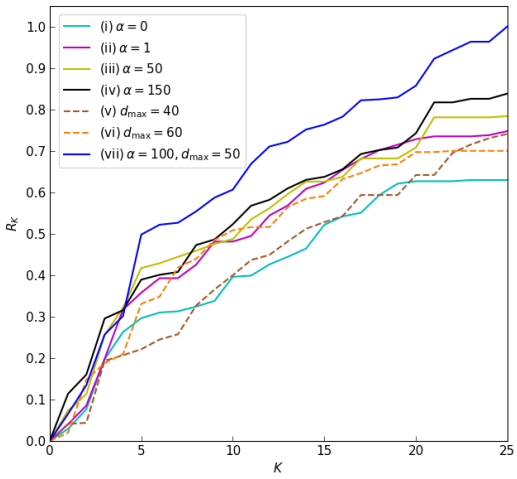


図 5: Sensitivity analysis の定量的結果.

次に，$f_{\mathrm{cov}}(A)$ が劣モジュラ性を持つことを以下に説明する．集合関数 $f$ が劣モジュラ性を持つとき $A \subseteq B$ を満たす任意の $A, B \subseteq E$ かつ任意の $e \subseteq E \backslash B$ について以下が成り立つ：

$$f_{\mathrm{cov}}(A \cup \{e\}) - f_{\mathrm{cov}}(A) \geq f_{\mathrm{cov}}(B \cup \{e\}) - f_{\mathrm{cov}}(B) \quad (1)$$

$f_{\mathrm{cov}}(A)$ は非負であり，$f_{\mathrm{cov}}(\varnothing) = 0$ であるとき，$f_{\mathrm{cov}}(A)$ は特定の条件下において劣モジュラ関数であることが証明される [Nemhauser 78]．したがって，式 (1) で示される $f_{\mathrm{cov}}(A)$ は劣モジュラ関数である．

$\boldsymbol{u}_{\mathrm{neg}}$ を導入しない場合，つまり，遮蔽物を考慮しない場合，SOPO は遮蔽物の反対側の値を含む観測姿勢を選択する可能性がある．この問題を対処するために，$\boldsymbol{u}_{\mathrm{obs}}$ が遮蔽物を含むとき，遮蔽物ペナルティ $\alpha \boldsymbol{u}_{\mathrm{neg}}$ を導入する．ここで，$\alpha$ は遮蔽物ペナルティの量を決定する遮蔽物パラメータを表す．

次の姿勢 $y$ が導入されたとき，$f_{\mathrm{cov}}(A)$ の増加量を $\delta_y$ とすると，$\delta_y = f_{\mathrm{cov}}(A \cup y) - f_{\mathrm{cov}}(A)$ である．また，$A$ が与えられたとき、次の観測姿勢を以下のように選択する．

$$\hat{y} = \underset{y \in V \backslash A}{\operatorname{argmax}} \delta_y \quad (2)$$

式 (2) を $K$ 回実行することで，観測姿勢集合 $A_K$ が得られる．

前述したように，COPO は典型的な NP 困難な組合せ最適化問題である．一方で，$f_{\mathrm{cov}}(A)$ が劣モジュラ関数であることを考慮すると，現実的な制約時間内に貪欲法によって近似的な最適姿勢群を選択可能である．実際に，貪欲法は常に最適解が得られるとは限らないものの，単純な劣モジュラ関数に対しては $(1 - 1/e)$ 近似が保証されている [Nemhauser 78]．これは，最悪のケースにおいて最適スコアの約 63%の達成を保証していることを示している．

# 4. 実験

## 4.1 実験設定

本研究では，実際の屋内環境で収集された標準データセットである Matterport3D [Chang 17] を基に，Gazebo を使用してタスク環境を構築した．環境は全て Matterport3D から取得され，検証環境として，Matterport3D に含まれる家庭環境の中から，10 個以上の日常物体を含む環境をランダムに選択した．本研究では，実環境から得られた 3D メッシュモデルの使用および実際のロボットを再現したモデルの使用という 2 つの条件下において実世界に近い環境を構築した．

本研究でのタスク環境における観測は連続空間で行うことが重要なため，Matterport3D データセットに含まれる 3D モデルを利用した Gazebo ベースのタスク環境を新たに構築した．また，ロボットは階を移動せず，各環境は 1 つの階のみを含むことを前提とする．5 種類のタスク環境には，平均して 6 種類の部屋，35.2 種類の家具，39.8 種類の日常物体が存在する．

シミュレータでは，トヨタ自動車製の Human Suppoprt Robot (HSR) [Yamamoto 19] を使用した．HSR は World Robot Summit 2020 Partner Robot Challenge/Real Space [WRS 20] における標準 DSR である．複数の環境において複数のプロンプトを比較し，最も多くの物体を検出したプロンプトを採用した．また，OMM では，事前に学習された OpenScene モデルを使用した．以下では，提案手法の入力である 2D マップの構築手順について説明する．まず，ROS 内の HectorSLAM [Kohlbrecher 11] モジュールを使用して 2D マップを構築した．次に，通路が狭いことにより HSR が移動できない領域を除外するために，2D マップを修正した．

評価尺度は，収集された画像群において観測された日常物体の割合 $R_K = \frac{1}{n_{\max}} \sum_{k=1}^{K} n_k$ を使用した．ここに，$n_k$ および $n_{\max}$ はそれぞれ第 $k$ 番目の観測姿勢において観測された日常物体数および環境内の日常物体数を示す．$n_k$ は Detic [Zhou 22] のような物体検出器を使用して得られた．また，適切に検出された物体のみを手動で数え，同じ物体が異なる姿勢における画像に含まれた場合，1 回のみ数に含んだ．本研究では，日常物体の効率的な観測を，限られた観測回数下において物体のカバレッジを最大化することと定義する．

## 4.2 定量的結果

ベースライン手法および提案手法の定量的結果を図 4 に示す．横軸および縦軸は，それぞれ $K$ および $R_K$ を表す．なお，実験は 5 環境において行い，その平均値および標準偏差を示す．ベースライン手法は観測姿勢集合をランダムに生成する random baseline および GPT-3.5 [gpt] によって観測姿勢の系列を生成する GPT-3.5 手法の 2 手法とした．GPT-3.5 手法はプロンプトに基づき 25 個の家具の名前のリストを生成する．観測姿勢は指定された家具の正面を向くように設定した．これらの手法を選択した理由は以下である．random baseline については，能動学習における典型的なベースラインとするためである．GPT-3.5 手法については，GPT-3.5 がナビゲーションを含む行動系列生成タスク [Biggie 23, Zhao 23] について良好な結果が報告されており，これが COPO に類似しているためである．

図 4 より，$K = 5$ の場合，提案手法の $R_5$ が 0.50 である一方で，random baseline および GPT-3.5 手法の $R_5$ はそれぞ

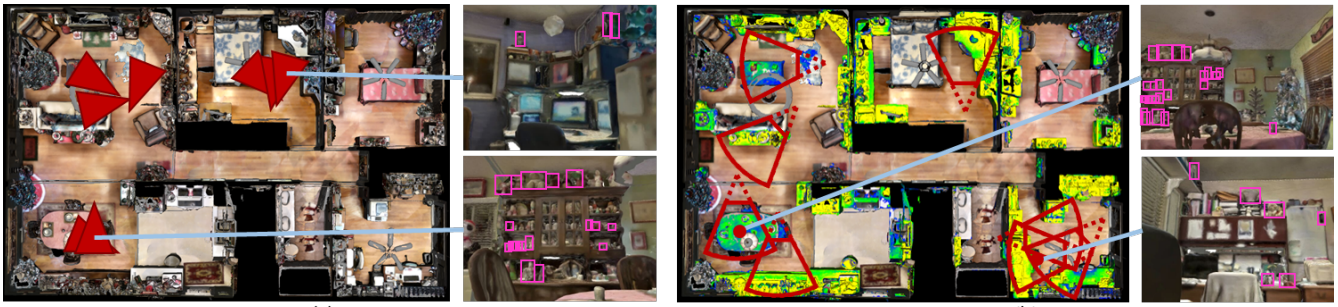


図 6: (a) GPT-3.5 手法および (b) SOPO による定性的結果.

れ 0.26 および 0.21 だった．したがって，提案手法は random baseline および GPT-3.5 手法をそれぞれ 0.24 および 0.29 ポイント上回った．これは，SOPO がさまざまな環境において効率的な観測姿勢を選択したことを示す．同様に，$K = 25$ の場合，提案手法，random baseline，GPT-3.5 手法の $R_{25}$ はそれぞれ 1.0，0.59，0.48 であった．この結果から，提案手法は観測姿勢の数が増加しても，random baseline および GPT-3.5 手法をそれぞれ 0.41 および 0.52 ポイント上回った．これは，観測姿勢数が増えても SOPO が効果的な観測姿勢を選択できたことを示す．

### 4.3 定性的結果

図 6 に (a)GPT-3.5 手法および (b)SOPO による定性的結果を示す．中央の画像は GPT-3.5 手法および SOPO によって得られた観測姿勢を示す．(b) では SOPO による positive object occurrence map も示す．(a) および (b) の右の 2 枚の画像は選択された観測姿勢において取得されたカメラ画像を示す．ピンクの四角形は検出された日常物体を表す．

$K = 8$ の場合，GPT-3.5 手法によって 39 個の日常物体が観測された一方，SOPO によって 59 個の日常物体が観測された．この環境における $n_{\max}$ は 86 であるため，それぞれの $R_8$ は 0.45 および 0.69 であった．これは，ロボットが効率的に日常物体を観測可能な観測姿勢を SOPO が選択したことを示す．

### 4.4 Sensitivity Analysis

図 5 に感度解析の定量的結果を示す．次の 2 つのパラメータを調査した．

**Selective wall parameter sensitivity investigation**

SPOM において，どのような値が効果的であるかを調査するために $\alpha$ を変更した．Model (i), (ii), (iii), (iv), および (vii) の $\alpha$ をそれぞれ 0, 1, 50, 150, および 100 に設定し，すべてのモデルの半径を 50 に設定した．図 5 より，$K = 25$ の場合，Model (vii) の $R_{25}$ は 1.0 であったが，Model (i), (ii), (iii) および (iv) の $R_{25}$ はそれぞれ 0.63, 0.75, 0.78, および 0.84 であった．これは，遮蔽物ペナルティが遮蔽物の考慮および効果的な最適化に貢献し，特に $\alpha = 100$ のモデルが貢献したことを示す．

**Radius sensitivity investigation**

$d_{\max}$ に対する感度を調査するために，SPOM において $d_{\max}$ を変更した．Model (v), (vi), および (vii) の $d_{\max}$ をそれぞれ 40, 60 および 50 に設定し，$\alpha$ はすべて 100 に設定した．図 5 より，$K = 25$ の場合，Model (v), (vi), および (vii) の $R_25$ はそれぞれ 1.0, 0.74, および 0.70 だった．これは，我々の手法が様々なカメラモデルで観測姿勢集合を最適化することが出来ることを示している．

## 5. おわりに

本研究では，組合せ最適化問題である NP 困難な COPO に着目し，観測可能な物体数を最大にするために，モデルが環境観測姿勢集合を選択した．提案手法は $R_K$ の指標でベースライン手法を上回り，ロボットが効率的に日常物体を観測することを可能にする観測姿勢を選択した．

**謝辞**

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

## 参考文献

[Biggie 23] Biggie, H., Mopidevi, A. N., Woods, D., et al.: Tell Me Where to Go: A Composable Framework for Context-Aware Embodied Robot Navigation, in *CoRL* (2023)

[Chang 17] Chang, A., Dai, A., Funkhouser, T., et al.: Matterport3D: Learning from RGB-D Data in Indoor Environments, in *3DV*, pp. 667–676 (2017)

[Choy 19] Choy, C., et al.: 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks, in *CVPR*, pp. 3075–3084 (2019)

[Ghiasi 22] Ghiasi, G., Gu, X., et al.: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels, in *ECCV*, pp. 540–557Springer (2022)

[gpt] https://platform.openai.com/docs/models/gpt-3-5

[Kaneda 24] Kaneda, K., et al.: Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine, *IEEE RA-L*, Vol. 9, No. 3, pp. 2088–2095 (2024)

[Kohlbrecher 11] Kohlbrecher, S., Von, O., et al.: A flexible and scalable SLAM system with full 3D motion estimation, in *SSRR*, pp. 155–160 (2011)

[Nemhauser 78] Nemhauser, G. L., et al.: An analysis of approximations for maximizing submodular set functions—I, *Mathematical programming*, Vol. 14, pp. 265–294 (1978)

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Sugiura 18] Sugiura, K.: SuMo-SS: Submodular Optimization Sensor Scattering for Deploying Sensor Networks by Drones, *IEEE RA-L*, Vol. 3, No. 4, pp. 2963–2970 (2018)

[WRS 20] World Robot Summit 2020 Partner robot challenge Real Space Rules & Regulations (2020)

[Yamamoto 19] Yamamoto, T., et al.: Development of Human Support Robot as the research platform of a domestic mobile manipulator, *ROBOMECH*, Vol. 6, No. 1, pp. 1–15 (2019)

[Zhao 23] Zhao, X., Li, M., Weber, C., Hafez, B., et al.: Chat with the Environment: Interactive Multimodal Perception Using Large Language Models, in *IROS* (2023)

[Zhou 22] Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I.: Detecting Twenty-thousand Classes using Image-level Supervision, in *ECCV*, pp. 350–368 (2022)