3D 視覚言語基盤モデルと劣モジュラ最適化による移動ロボットの環境探索

○鈴木駿太郎，松尾榛夏，杉浦孔明（慶應義塾大学）

本研究では，屋内環境の効率的な観測を目的とする，ロボットの観測姿勢集合における組合せ最適化問題を扱う．本タスクでは，できる限り多くの日常物体が観測可能であることが望ましい．これは，NP 困難な問題であり，観測姿勢数の増加に伴い組合せ爆発を起こす．既存手法では，観測時の遮蔽について遮蔽物体の存在領域のみを考慮しており，視野の遮蔽についての考慮が不十分であった．そこで本研究では，遮蔽を考慮した，観測姿勢集合の劣モジュラ最適化手法 Occlusion Aware SOPO (OA-SOPO) を提案する．本手法の評価に際して，Matterport3D を基盤とするタスク環境を構築し，選択された観測姿勢集合を用いて観測画像を収集した．結果として，OA-SOPO は環境内における日常物体の観測割合においてベースライン手法を 0.32 ポイント上回った．

1. はじめに

移動ロボットは，災害救助，生活支援，アクセス困難地域での監視，物資の配送，農業での収穫作業など，多岐にわたる分野での活躍が期待されている．特に，生活支援ロボットの活用においては，高齢者や障がいを持つ人々の，家の中での日常生活におけるタスクを効果的に支援する事が期待される．こうしたロボットにとって，屋内環境における環境情報を事前に把握することは，日常タスクを効果的に支援する上で重要である．

そこで，本研究では Combinatorial Observation Pose Optimization (COPO) タスクに着目する．COPO タスクは，効率的な環境観測のためのロボット姿勢集合の組合せ最適化問題であり，NP 困難である．観測姿勢数の増加に伴い観測姿勢集合の選択における計算量は指数関数的に増加する．しかし，無数の観測姿勢を全て網羅することは現実的でない．よって，環境に含まれる日常物体を効率的に観測する観測姿勢集合の選択は困難である．また，各観測姿勢において観測可能な日常物体は物体間の遮蔽に影響されるため，観測姿勢の選択時には観測領域の遮蔽の考慮が求められる．

本タスクにおいて良好な結果が報告されている SOPO [1] では，劣モジュラ性および open-vocabulary な 3D 特徴量を用いて生成された物体存在マップ群を用いて観測姿勢集合を選択している．しかし，環境と物体存在マップ間の一部に乖離が報告されている．この乖離に起因して，選択された観測姿勢集合により取得された画像群の一部において，日常物体が配置されていない領域を観測してしまう傾向があった．また，観測時の遮蔽については，遮蔽物体の存在領域のみを考慮しており，視野の遮蔽についての考慮は不十分であった．そのため，壁により観測が妨げられてしまう傾向があった．

本研究では，Occlusion Aware SOPO (OA-SOPO) を提案する．これは，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り，屋内環境においてできる限り多くの日常物体の観測を目的とする，OA-SOPO では，open-vocaburaly な 3D 特徴量およびテキストプロンプト群から生成した 3 種の物体存在マップを用いて，各観測姿勢により変化する観測領域の物体存在スコアを評価する．加えて，観測姿勢における遮蔽を考慮するための Adaptive Obejct Occurrence Scorer (AOOS) を導入する．

提案手法における新規性は以下である．

- ロボットの観測領域を考慮し，屋内環境を効率的に観測するための観測姿勢集合を選択する OA-SOPO を提案する．
- 各観測姿勢における物体存在スコアを評価するため，Path Free Map により物体存在領域を限定し，Positive Occurrence Map および Negative Occurrence Map によりロボットの観測領域における遮蔽を考慮する AOOS を導入する．

図 1　COPO の代表例

2. 関連研究

Active Sensing の分野では，多くの既存研究が存在する [1–4]．ロボットの coverage タスクを扱う [3,4] では，情報利得の最大化を目的として，グラフマッチングおよび探索ベースの手法を用いた経路の最適化手法を提案している．[1] は，open-vocabulary な 3D 特徴量および劣モジュラ最適化手法を用いた，ロボット姿勢集合の選択手法を提案している．本タスクでは，単数 [1,3] および複数 [4] のロボットが環境内を複数視点から観測する状況において，それぞれの視点から取得された情報間の影響を考慮することが求められる．

3. 問題設定

本研究では，ロボットの 2D 姿勢集合最適化を目的とする COPO タスクを扱う．これは，典型的な組合せ最適化問題であり，NP 困難である．ここで，最適化された姿勢集合とは環境内の物体を効率的に観測するロボット姿勢の組合せを表す．図 1 に本タスクにおける代表例を示す．

本タスクにおける入力は，事前に探索して得られた屋内環境の 2D マップおよび環境中の家具に関する 3D 点群である．出力は，環境内の物体観測数を最大化する観測姿勢の集合である．本論文では，日常物体を，生活支援ロボットによる把持や移動の対象となる物体と定義する．具体例として，本やコップが挙げられる．また，遮蔽物体について，一定以上の高さがあり，生活

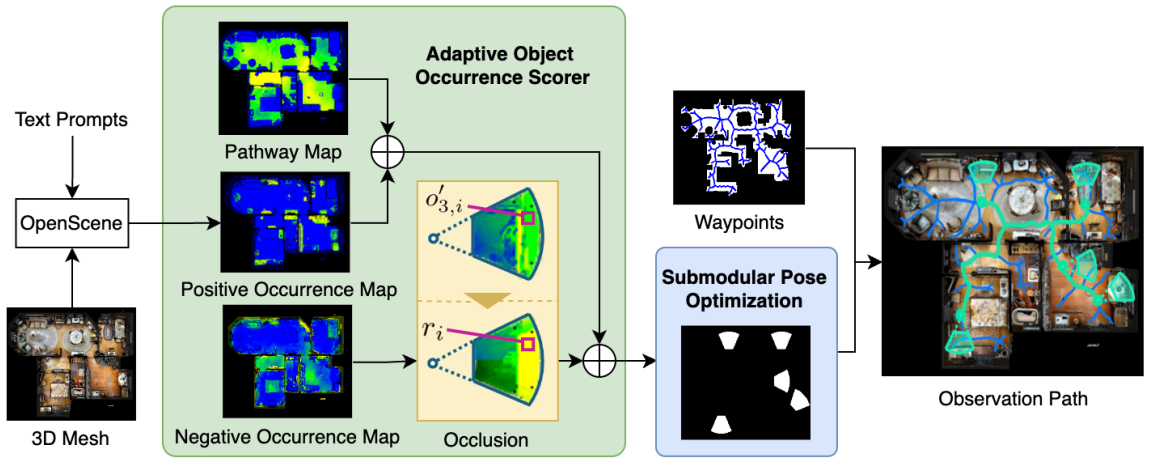

図 2　提案手法のモデル構造

支援ロボットの物体探索において視野を妨げる物体と定義する．具体例として，本棚や壁が挙げられる．また，観測姿勢を屋内環境の 2D マップ上における座標及び向きと定義する．また，本研究では，日常物体の観測を主眼とし，その操作は扱わない．そして，環境観測に用いるカメラは 1 台に限定し，そのパラメータは既知とする．

4.　手法

図 2 に提案手法のモデル構造を示す．提案手法は 2 種類の主要モジュールで構成されており，AOOS および Submodular Pose Optimization Module (SPOpt) である．本モデルへの入力は 2 種類である．1 つ目は，環境内の家具に関する点群 $\boldsymbol{x}_{\text{pcl}} \in \mathbb{R}^{M\times 3}$ 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 $\boldsymbol{x}_{\text{map}} \in \mathbb{R}^{H\times W}$ であり，H，W はそれぞれ 2D 占有格子地図の行数および列数を表す．

4.1　Adaptive Object Occurrence Scorer

AOOS は，各観測姿勢の観測領域に対し，日常物体の存在スコアを評価する．物体存在スコアの導出に際して，open-vocabulary な 3D 特徴量に基づく，物体存在マップを活用する．これにより，大規模データセットに基づく常識的な推論が可能となり，多様な物体の存在スコアを評価可能となる．ただし，open-vocabulary な 3D 特徴量 $\boldsymbol{F} \in \mathbb{R}^{M\times C}$ は SOPO [5] と同様，OpenScene [5] により取得する．ここで，C はセグメンテーションモデルにより出力される特徴次元数を表す．

物体存在マップを得るために，$\boldsymbol{F}$ に対して 3 種のテキストプロンプト P_1，P_2，P_3 をそれぞれ使用し，それぞれのプロンプトに対して物体存在スコアの高い領域 $\boldsymbol{F}_{\text{likelihood}} \in \mathbb{R}^{3\times M\times C}$ を取得する．ここで，3 種のプロンプトは，以下を使用した．

- P_1: “Places that are pathway”
- P_2: “Places to put objects that can be carried”
- P_3: “Wall and places that occlude vision”

P_1 は通路の領域を取得するためのプロンプトである．環境内の通路には日常物体が配置されていないという仮定の下，観測姿勢がカバーすべき観測領域を限定する意図がある．また，P_2 は棚や机といった日常物体が置かれやすい領域を取得するためのプロンプトである．

Algorithm 1 各観測姿勢の物体存在スコアの導出

Input: $\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3, \boldsymbol{a}'$
Output: $\boldsymbol{o}_{\boldsymbol{a}'}$
$(a'_x, a'_y) = \boldsymbol{a}'$
$\boldsymbol{a}'_{\text{cov}} = C(\boldsymbol{a}')$
$\boldsymbol{q} = \boldsymbol{1} - \boldsymbol{o}_1$
$(\boldsymbol{q}', \boldsymbol{o}'_2, \boldsymbol{o}'_3) = \boldsymbol{a}'_{\text{cov}} \odot (\boldsymbol{q}, \boldsymbol{o}_2, \boldsymbol{o}_3)$
for $i = 1$ *to* $H \times W$ **do**
    $r_i = \text{Occlusion}(\text{a}'_{\text{x}},\ \text{a}'_{\text{y}},\ \text{o}'_{3,\text{i}})$
    if $q'_i > 0$ *and* $o'_{2,i} > 0$ **then**
        $o_{\boldsymbol{a}',i} = \alpha\ (q'_i + o'_{2,i})$
    else
        $o_{\boldsymbol{a}',i} = q'_i + o'_{2,i}$
    end if
    $o_{\boldsymbol{a}',i} = (o_{\boldsymbol{a}',i} - \beta \cdot r_i).\text{Clip}(0, \text{inf})$
end for
$\boldsymbol{o}_{\boldsymbol{a}'} := \{o_{\boldsymbol{a}',i} | i = 1, ..., H \times W\} \in \mathbb{R}^{H\times W}$

そして，P_3 は遮蔽物体の存在する領域を取得するためのプロンプトである．これにより，ロボットの視野の遮蔽を考慮する．ただし，それぞれのプロンプトの組合せを事前に比較し，最も多くの日常物体の観測に寄与した組合せを採用した．

そして，得られた点群特徴量をそれぞれ 2D グレースケール画像に変換し，Path Free Map $\boldsymbol{o}_1 = \{o_{1,i} | i = 1,\ \ldots\ , H \times W\}$ および Positive Occurrence Map $\boldsymbol{o}_2 = \{o_{2,i} | i = 1,\ \ldots\ , H \times W\}$，Negative Occurence Map $\boldsymbol{o}_3 = \{o_{3,i} | i = 1,\ \ldots\ , H \times W\}$ をそれぞれ得る．ただし，グレースケール画像の各ピクセルは閉区間 [0,1] に正規化されている．

次に，各観測姿勢における観測領域を考える．カメラモデルは SOPO と同様のモデルを定義する．このとき，ある観測姿勢 $\boldsymbol{a}'$ がカバーする観測領域 $C(\boldsymbol{a}')$ は以下で定式化される．

$$\boldsymbol{a}'_{\text{cov}} = C(\boldsymbol{a}') = \begin{cases} 1 \text{ if coverage area} \\ 0 \text{ otherwise} \end{cases}$$

Algorithm 1 に $\boldsymbol{a}'$ の観測領域における物体存在スコア $\boldsymbol{o}_{\boldsymbol{a}'} \in \mathbb{R}^{H\times W}$ の導出アルゴリズムを示す．アルゴリズムにおける入力および出力は，$\{\boldsymbol{o}_1,\ \boldsymbol{o}_2,\ \boldsymbol{o}_3,\ \boldsymbol{a}'\}$ および $\boldsymbol{o}_{\boldsymbol{a}'} \in \mathbb{R}^{H\times W}$ である．初めに，$\boldsymbol{a}'_{\text{cov}}$ および x 座標 a'_x，y 座標 a'_y をそれぞれ取得する．次に，$\boldsymbol{o}_1$ で表されるグレースケール画像を反転し，通路以外の領域 $\boldsymbol{q}$ を特

定する．これは，通路以外の領域を取得することで，観測姿勢がカバーすべき観測領域を限定する意図がある．そして，$\boldsymbol{o}_1$，$\boldsymbol{o}_2$，$\boldsymbol{o}_3$ をそれぞれ $\boldsymbol{a}'_{\mathrm{cov}}$ でマスクすることにより，カメラモデルのカバー領域を考慮した物体存在マップ $\boldsymbol{q}' \in \mathbb{R}^{H\times W}$，$\boldsymbol{o}'_2 \in \mathbb{R}^{H\times W}$，$\boldsymbol{o}'_3 \in \mathbb{R}^{H\times W}$ をそれぞれ得る．その後，$\boldsymbol{q}' \in \mathbb{R}^{H\times W}$ および $\boldsymbol{o}'_2 \in \mathbb{R}^{H\times W}$，$\boldsymbol{o}'_3 \in \mathbb{R}^{H\times W}$ における i 番目の微小領域 q'_i および $o'_{2,i}$，$o'_{3,i}$ を以下の処理により集約する．

初めに，$\boldsymbol{a}'$ から $o'_{3,i}$ を観測する際の遮蔽を考慮するための操作として，$\mathrm{Occlusion}(a'_x, a'_y, o'_{3,i})$ を定義する．Occlusion では，(a'_x, a'_y) および $o'_{3,i}$ を入力として，両者の線分を考える．そして，線分上におけるスコアの最大値 r_i を出力とする．ここでは，$\boldsymbol{a}'$ からの観測において $o'_{3,i}$ を遮蔽する領域のスコアを新たに取得している．

次に，q'_i および $o'_{2,i}$ を集約する．このとき，q'_i および $o'_{2,i}$ の両者において正のスコアが得られた場合は係数 α により重み付けする．これは，重み付けの対象である微小観測領域およびその周辺から複数の観測姿勢が選択されることを意図する．これにより，日常物体が存在する可能性の高い領域を確実に観測することを可能にする．最後に，q'_1 および $o'_{2,i}$ の集約項に r_i を集約し，$\boldsymbol{a}'$ における物体存在スコア $o_{\boldsymbol{a}',i}$ を得る．ここで，r_i の集約では正規化係数 β を設ける．これは，係数 α により重み付けされた物体存在スコアと r_i を調整する項である．

以上の処理を $\boldsymbol{q}'$ および $\boldsymbol{o}'_2$，$\boldsymbol{o}'_3$ に関する全ての微小観測領域に対して行うことで，$\boldsymbol{a}'$ の物体存在マップ $\boldsymbol{o}_{\boldsymbol{a}'} = \{o_{a',i} | i = 1, \ldots, H \times W\} \in \mathbb{R}^{H\times W}$ を得る．このようにして，各観測姿勢における物体存在スコアの集合が出力 $\boldsymbol{o}_V = \{o_{\boldsymbol{a}'} | \boldsymbol{a}' \in V\}$ として得られる．ここで，V は入力 $\boldsymbol{x}_{\mathrm{map}}$ から得られる観測姿勢の候補集合を表す．

4.2 Submodular Pose Optimization

SPOpt は，カバレッジを最大化する観測姿勢集合を選択するモジュールであり，SOPO と同様の構造である．ここでは，劣モジュラ性に基づく貪欲法 [6] を利用することで，現実的な制約時間内に近似的な最適姿勢集合を選択する．本モジュールにおける入力は $\boldsymbol{o}_V$ および $\boldsymbol{x}_{\mathrm{map}}$ であり，出力は観測姿勢集合 A_K である．ただし，K は集合のサイズを表す．

5. 実験設定

本研究では，実環境から取得された標準的なデータセットである Matterport3D [7] に基づいて，Gazebo シミュレータ上でタスク環境を 10 種構築した．検証環境は Matterport3D に含まれる屋内環境のうち，日常物体が 10 個以上配置された環境をランダムに選択した．本研究では，連続的な空間における観測が求められるため，DSR による連続的な観測点への移動が可能なタスク環境を新たに構築した．ただし，DSR は階の移動を想定しておらず，各構築環境は 1 階分のみを含むものとする．ここで，DSR はトヨタ自動車製の Human Support Robot（HSR）[8] を使用した．HSR は World Robot Summit 2020 Partner Robot Challenge/Real Space における標準 DSR である．また，検証における 10 種の屋内環境は，平均して 6.4 部屋および 31.4 個の家具，36.1 個の日常物体を含む．

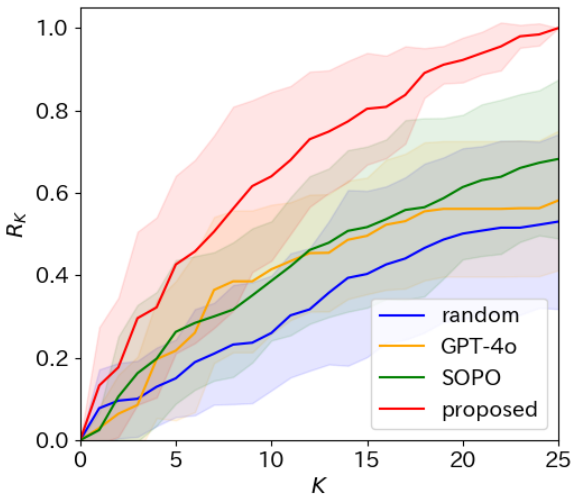

図 3　提案手法およびベースライン手法の比較

次に，本タスクにおける 2D マップの構築手順を示す．初めに，HSR に搭載された LiDAR センサを用い，ROS の HectorSLAM [9] モジュールから 2D Map を構築した．その後，生成した 2D マップを用いて実際に HSR による巡回を実施し，HSR の通過不能領域を特定した．これらの通過不能領域を対象外とするために 2D マップを修正し，提案手法の入力とした．加えて，観測姿勢集合の巡回では，2D マップからボロノイ図を作成し，巡回セールスマン問題を解くことで移動経路を求めた．

6. 実験結果

6.1 定量的結果

提案手法およびベースライン手法の定量的結果を図 3 に示す．横軸と縦軸はそれぞれ K および環境内の日常物体総数に対する観測できた日常物体数の割合を表す．なお，実験は 10 環境で行い，その平均値および標準偏差を示す．ベースライン手法として，SOPO [1]，Random Method，GPT-4o Method の 3 手法を選択した．Random Method ではランダムな観測姿勢集合を選択した．GPT-4o Method では GPT-4o [10] を用いてプロンプトから観測姿勢集合を選択した．これらの手法を選択した理由は以下である．SOPO は，COPO において良好な結果が報告されているため選択した．Random Method は，能動学習における代表的なベースライン手法とするため選択した．そして，GPT-4o Method はナビゲーションを含む行動系列生成タスク [11] において良好な結果が報告されているため選択した．

また，評価尺度は，環境の日常物体総数に対する，観測姿勢集合からの収集画像群において観測された日常物体数の割合 R_K を用いた．ここで，R_K は $R_K = \frac{1}{n_{\max}} \sum_{k=1}^{K} n_k$ で定義する．ただし，n_k および $n_{\max}$ はそれぞれ k 番目の観測姿勢から観測された日常物体数および環境内の日常物体総数を表す．なお n_k は，Detic [12] に代表される標準的な物体検出器を使用し，適切に検出された日常物体のみを手動で数えることにより得た．このとき，異なる観測姿勢から検出された同一物体は一度のみ n_k に含めた．本研究では，効率的な日常物体の観測を，限られた観測姿勢数における日常物体のカバレッジ最大化と定義する．

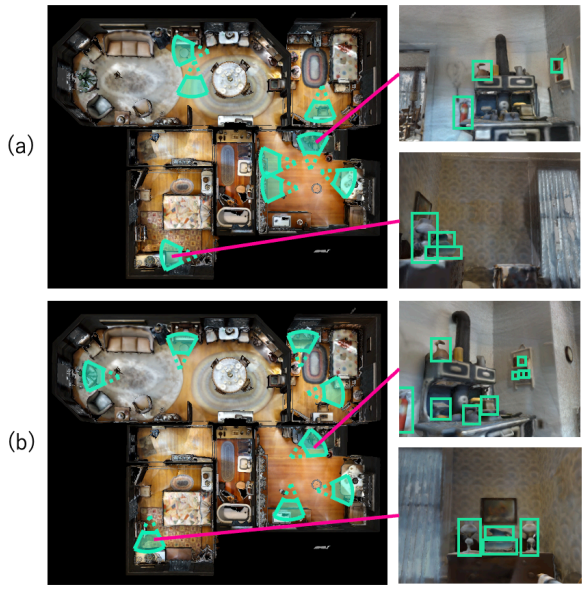

図 4　(a)SOPO および (b) 提案手法による定性的結果

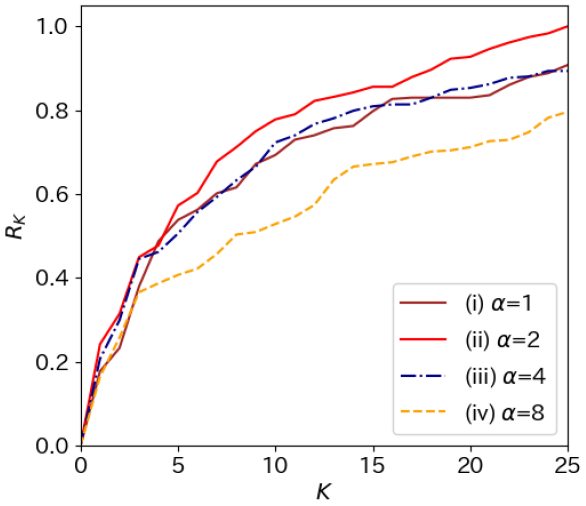

図 5　感度解析の定量的結果

図 3 より K=5 の場合，SOPO，Random Method，GPT-4o Method，提案手法における R_5 は 10 環境で平均してそれぞれ 0.26，0.15，0.22，0.43 であった．したがって，提案手法は SOPO，Random Method，GPT-4o Method をそれぞれ 0.17，0.28，0.21 上回っており，提案手法が多様な環境で効率的な観測姿勢集合を選択可能であると示唆される．同様に，K=25 の場合，SOPO，Random Method，GPT-4o Method，提案手法における R_{25} は 10 環境で平均してそれぞれ 0.68，0.53，0.58，1.0 であった．したがって，提案手法は SOPO，Random Method，GPT-4o Method をそれぞれ 0.32，0.47，0.42 上回っており，提案手法では観測姿勢の数が増加した場合でも環境内の物体を多く観測する観測姿勢集合が選択可能であると示唆される．

6.2　定性的結果

図 4 に，K=8 における（a）提案手法および（b）SOPO の定性的結果を示す．ここに，左の画像は提案手法および SOPO を用いて選択された A_8 を示している．また，右の画像は A_8 において収集された観測画像 2 例を示しており，緑色の矩形領域は適切に検出された日常物体を示している．ここで，図 4 において提案手法および SOPO の選択姿勢集合から収集された観測画像は，それぞれ環境内の同じ領域を異なる姿勢から観測した．このとき，提案手法および SOPO の 2 例の観測画像において，適切に検出された日常物体の総数は，それぞれ 6 個および 13 個であった．したがって，提案手法では観測範囲内の各点における遮蔽を考慮したスコアを用いることにより，効率的な観測姿勢集合の選択が可能であると示唆される．

6.3　感度解析

図 5 に感度解析の定量的結果を示す．感度解析では α に関して効果的な値を調査した．なお，実験は 5 環境で行い，その平均値を示す．本手法における α を (i)，(ii)，(iii)，(iv) においてそれぞれ 1，2，4，8 と定めた．図 5 より，R_{25} は (i)，(ii)，(iii)，(iv) においてそれぞれ，0.91，1.0，0.89，0.80 であった．このことから，$\boldsymbol{o}_1$ および $\boldsymbol{o}_2$ 間における物体存在スコアの重複を考慮する重み付け項 α は，効果的な最適化に貢献したと示唆される．特に，$\alpha = 2$ のモデルが効果的であったことが示唆される．

7.　結論

本研究では，組合せ最適化問題であり NP 困難である COPO タスクに着目した．COPO タスクでは，モデルは観測可能な物体数を最大化するための環境観測姿勢集合を選択した．10 環境を用いて実験を行い，ベースライン手法を上回る結果が得られた．

謝辞

本研究の一部は，JSPS 科研費 23K28168，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

[1] 松尾榛夏，神原元就，杉浦孔明，“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，” 第 38 回人工知能学会全国大会，2024.
[2] K. Sugiura, “SuMo-SS: Submodular Optimization Sensor Scattering for Deploying Sensor Networks by Drones,” *IEEE RA-L*, vol.3, no.4, pp.2963–2970, 2018.
[3] T. Kusnur, D. Saxena, and M. Likhachev, “Search-Based Planning for Active Sensing in Goal-Directed Coverage Tasks,” in *ICRA*, pp.15–21, 2021.
[4] N. Karapetyan, A.B. Asghar, et al., “AG-CVG: Coverage Planning with a Mobile Recharging UGV and an Energy-Constrained UAV,” in *ICRA*, pp.2617–2623, 2024.
[5] S. Peng, et al., “OpenScene: 3D Scene Understanding with Open Vocabularies,” in *CVPR*, pp.815–824, 2023.
[6] G. Nemhauser, et al., “An Analysis of Approximations for Maximizing Submodular Set Functions—I,” *Mathematical Programming*, vol.14, pp.265–294, 1978.
[7] A. Chang, et al., “Matterport3D: Learning from RGB-D Data in Indoor Environments,” in *3DV*, pp.667–676, 2017.
[8] T. Yamamoto, 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.
[9] S. Kohlbrecher, O. Von, J. Meyer, and U. Klingauf, “A Flexible and Scalable SLAM System with Full 3D Motion estimation,” in *SSRR*, pp.155–160, 2011.
[10] OpenAI, “GPT-4o,” Accessed: June. 2024. Available: https://platform.openai.com/docs/models/gpt-4o
[11] H. Biggie, A. Mopidevi, D. Woods, et al., “Tell Me Where to Go: A Composable Framework for Context-Aware Embodied Robot Navigation,” in *CoRL*, 2023.
[12] X. Zhou, et al., “Detecting Twenty-thousand Classes using Image-level Supervision,” in *ECCV*, pp.350–368, 2022.