マルチモーダル基盤モデルと緩和対照損失を用いた大規模屋内検索エンジン

Large-Scale Indoor Search Engine with Multimodal Foundation Models and Relaxing Contrastive Loss

今井 悠人 Yuto Imai　兼田 寛大 Kanta Kaneda　是方 諒介 Ryosuke Korekata　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

In this paper, we focus on the learning-to-rank physical objects task. In this task, images of objects within large-scale indoor environments are ranked based on open-vocabulary user instructions. We introduce the GREP module to construct visual features considering image, target object, relative positions, and pixel granularities. Additionally, we introduce the RCS module to efficiently learn from redundant images taken in the indoor environment. Our method outperformed baseline methods on the newly constructed YAGAMI dataset and an extended LTRRIE-subset, showing significant improvements in the standard metrics.

1. はじめに

少子高齢化に伴う労働人口の不足は喫緊の社会問題となっている．この問題に対し，モバイルロボットは人的労力を削減できるという点で期待されている．実世界におけるモバイルロボットの応用のためのアプローチとして，老朽化の監視や安全管理などの異常検知，物体の運搬などの既知環境における定期的な移動を伴うタスクを自動化させることが考えられる．これらは安全性や頻度の面から，すべて人間が行う場合大きな人的労力が必要となる一方で，完全に自動化させるには未だ課題が多い．そこで，自動化とオペレータによる介入を組合せた human-in-the-loop 設定を導入することで，この両方の問題の解決が期待できる．

本研究では，人間がロボットに実環境中の物体に関する自然言語による指示を与えた時，対象物体を検索する Learning-to-rank physical objects（LTRPO）タスク [Kaneda 24] を扱う．LTRPO タスクは，既存の参照表現理解タスク [Korekata 23] と関連が深い．一方で，LTRPO タスクで扱う，複雑な参照表現を含む open-vocabulary な指示文は曖昧さや冗長性を含むことが多いため，指示文から対象物体を正しく把握することは難しい．

本論文では，大規模屋内空間中の検索を行う手法を提案する．既存手法と異なる点は，検索空間を部屋単位から公共性の高い大規模環境への拡張，画像単位で事前学習された基盤モデル内の中間特徴の特徴マップから空間的な特徴および画素単位での意味的な特徴の抽出および検索空間の拡張に伴う学習の効率化を行った点である．

2. 問題設定

本タスクはユーザによる物体に関する open-vocabulary な指示文から，モバイルロボットが屋内環境で撮影した物体の領域画像を検索するタスクである．図 1 に LTRPO タスクの具体例を示す．指示文として，“Check that the washing machine on the left is working.” が与えられた場合に，モデルは対象物体を上位にランク付けしたリストを出力することが望ましい．本論文で扱う用語を以下のように定義する．

連絡先: 今井悠人，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，ytim8812@keio.jp

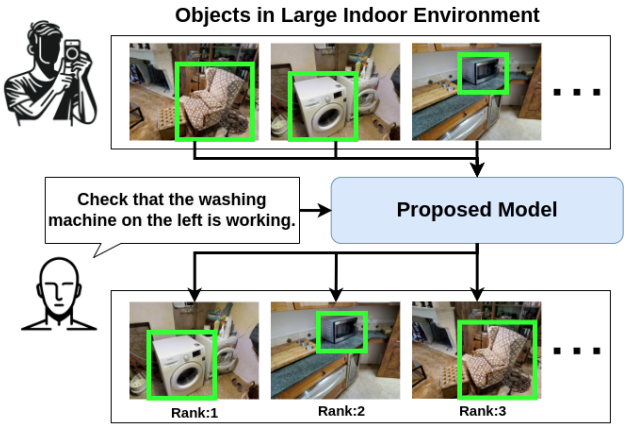

図 1: 本研究で扱う LTRPO タスクの例.

- **指示文:** 環境中の物体のもとに向かうために，モバイルロボットに与える文
- **対象物体:** 指示文の対象となる物体
- **対象物体領域:** 対象物体の矩形領域
- **環境画像:** 屋内で撮影された，対象物体が写っている画像

3. 提案手法

提案手法は Granular Representation from Entire to Pixels (GREP)，Crossmodal Noun Phrase Encoder (CNPE)，Relaxing Contrastive Similarity (RCS) の 3 つのモジュールから構成される．

3.1 入力

提案手法の入力 $\boldsymbol{x}$ は以下の形で定義される．

$$\boldsymbol{x} = (\boldsymbol{x}_{\text{inst}}, \mathcal{T}, \mathcal{C})$$

$$\mathcal{T} = \left\{ \boldsymbol{x}_t^{(n)} \middle| n = 1, ..., N_{\text{targ}} \right\}$$

$$\mathcal{C} = \left\{ \boldsymbol{x}_c^{(n)} \middle| n = 1, ..., N_{\text{targ}} \right\}$$

ここで，$\boldsymbol{x}_{\text{inst}} \in \{0,1\}^{V \times L}$，$\boldsymbol{x}_t^{(n)} \in \mathbb{R}^{3 \times H_t \times W_t}$，$\boldsymbol{x}_c^{(n)} \in \mathbb{R}^{3 \times H_c \times W_c}$ は one-hot ベクトルで表現される指示文，対象物体領域，環境画像である．またこの時，V，L，H_t，W_t，H_c，W_c はそれぞれ，語彙サイズ，最大トークン数，対象物体領域

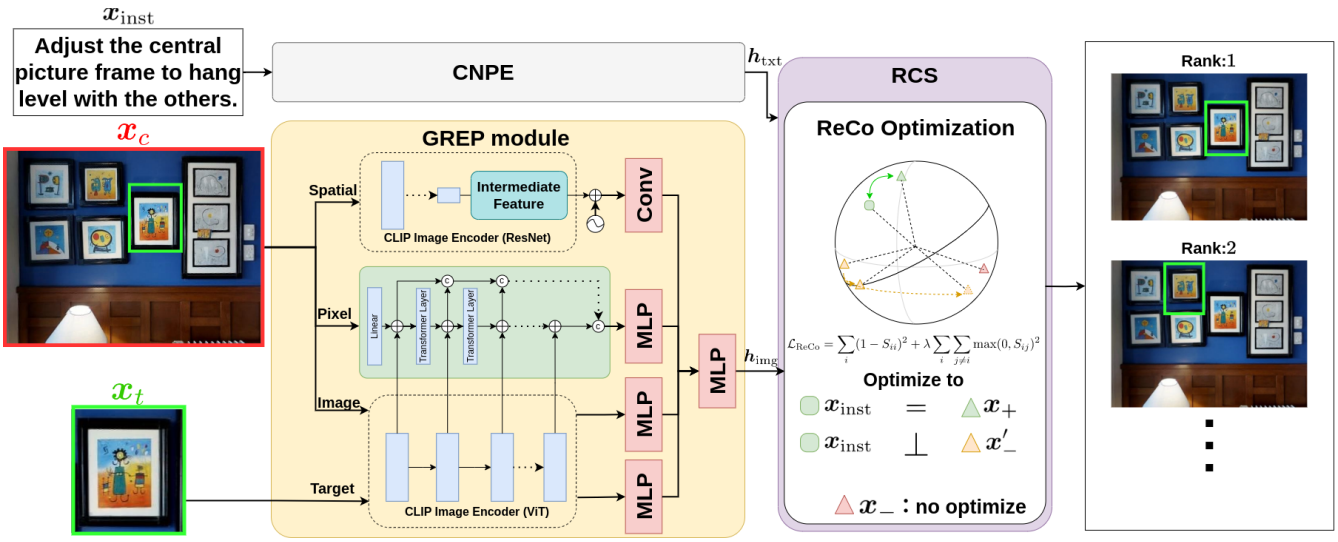

図 2: 提案手法のネットワーク構造

の高さ，幅，環境画像の高さ，幅を表す．

3.2 GREP モジュール

画像中の物体理解が重要となる LTRPO タスクでは，環境画像や対象物体領域から事前学習済みのエンコーダを通じて得られる特徴を単に組み合わせるだけでは十分ではない．例えば，$\boldsymbol{x}_{\text{inst}}$ が複雑な参照表現を含む場合，物体間の関係性を反映した接地が重要である．また，物体内部の特徴（構造，色，柄など）を参照表現として $\boldsymbol{x}_{\text{inst}}$ が含む場合，画素単位での意味的な物体理解が可能であることが望ましい．そこで本研究では，画像から様々な粒度で接地するための表現を扱う GREP モジュールを導入する．

GREP モジュールでは，$\boldsymbol{x}_t$ および $\boldsymbol{x}_c$ の組から，GREP を獲得し，また集約することで画像特徴量 $\boldsymbol{h}_{\text{img}}$ を抽出する．GREP は，Pixel-wise，Spatial-wise，Image-wise，Target-wise，の 4 つの粒度によって定義される．

Pixel-wise: 本タスクにおいて，$\boldsymbol{x}_c^{(n)}$ 全体に対象物体が映っているとは限らない．このような場合，CLIP が学習時に獲得している大域的な情報だけでは不十分であると考えられる．CLIP から局所的な特徴を得る方法として，CLIP を用いた open-vocabulary semantic segmentation モデルを適用することが考えられる．SAN [Xu 23] は CLIP を open-vocabulary semantic segmentation タスクに特化して学習されており，入力された単語に各画素が該当するかどうか画素単位の予測を行う．そこでここでは，SAN を参考として局所的で重要な画像特徴量を抽出する．まず，$\boldsymbol{x}_c^{(n)}$ を SAN に入力し，以下の式から中間特徴 $\boldsymbol{h}_{\text{SAN}}^{(i)}$ を得る．

$$\boldsymbol{h}_{\text{SAN}}^{(i)} = \text{CLIP}^{(i_c)}\left(\boldsymbol{x}_c^{(n)}\right) + \text{Transformer}^{(i_t)}\left(\boldsymbol{x}_c^{(n)}\right)$$

ここで，$\text{CLIP}^{(i_c)}$ および $\text{Transformer}^{(i_t)}$ は CLIP 画像エンコーダ (ViT-B/16) 第 i_c 層および第 i_t 層の transformer 層を表す．Pixel-wise 全体の出力は $\boldsymbol{h}_{\text{SAN}}^{(i)}$ を MLP に入力して得られる $\boldsymbol{h}_{\text{px}}$ である．

Spatial-wise: 上述したように，物体間の関係性や相対的な位置関係を抽出するために，2 次元の中間特徴を用いる．$\boldsymbol{x}_c^{(n)}$ に対し，CLIP 画像エンコーダ (RN50x4) の中間層出力に Positional Encoding [Vaswani 17] を加えたものを $\boldsymbol{h}_{\text{mid}}$ とする．この $\boldsymbol{h}_{\text{mid}}$ に対し畳み込みを繰り返すことで，空間的な画像特徴 $\boldsymbol{h}_{\text{sp}}$ を得る．

Image-wise: $\boldsymbol{x}_c^{(n)}$ から，事前学習済みの CLIP 画像エンコーダ (ViT-L/14) を用いて，画像全体に関する画像特徴 $\boldsymbol{h}_c^{(n)}$ を抽出する．

Target-wise: Image-wise と同様に，$\boldsymbol{x}_t^{(n)}$ に対し，CLIP 画像エンコーダ (ViT-L/14) [Radford 21] を適用して物体単位での画像特徴 $\boldsymbol{h}_t^{(n)}$ を抽出する．

GREP モジュールの出力は，各粒度の特徴を組合せて得られる画像特徴 $\boldsymbol{h}_{\text{img}}$ であり，以下の式から得られる．

$$\boldsymbol{h}_{\text{img}} = \text{MLP}\left([\boldsymbol{h}_{\text{c}}; \boldsymbol{h}_{\text{t}}; \boldsymbol{h}_{\text{sp}}; \boldsymbol{h}_{\text{px}}]\right)$$

3.3 CNPE

指示文中の対象物体が含まれる名詞句，その他の名詞句および前置詞句との関係を効果的にモデル化するために，CNPE モジュール [Kaneda 24] を導入する．まず，$\boldsymbol{x}_{\text{inst}}$ に [Schuster 16] を適用し，N_{noun} 個の名詞句 $\boldsymbol{x}_{\text{noun}}^{(k)}(k = 1, ..., N_{\text{noun}})$ を得る．その後，CLIP テキストエンコーダ [Radford 21] を用いて $\boldsymbol{x}_{\text{inst}}$ および $\boldsymbol{x}_{\text{noun}}^{(k)}$ から指示文に関する言語特徴量 $\boldsymbol{h}_{\text{inst}}$ および $\boldsymbol{h}_{\text{noun}}^{(k)}$ を獲得する．最後に，transformer 層に $\boldsymbol{h}_{\text{noun}}^{(k)}$ を連結したものを入力し，この出力と $\boldsymbol{h}_{\text{inst}}$ を MLP に入力することでモジュールの出力となる言語特徴量 $\boldsymbol{h}_{\text{txt}}$ が得られる．

3.4 RCS

RCS では，CNPE と GREP から得られた $\boldsymbol{h}_{\text{txt}}$ および $\boldsymbol{h}_{\text{img}}$ によって得られるコサイン類似度 s を出力する．s は，以下の式で計算される．

$$s(\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_t^{(n)}) = \frac{\boldsymbol{h}_{\text{txt}} \cdot \boldsymbol{h}_{\text{img}}}{||\boldsymbol{h}_{\text{txt}}||\ ||\boldsymbol{h}_{\text{img}}||}$$

モデルの出力は s に基づきランク付けされた画像集合 $\mathcal{T}$’ である．

また，ここで計算した s に基づき，損失を計算する．本研究では損失関数として，Relaxed Contrastive (ReCo) 損失 [Lin 23] および InfoNCE 損失 [Oord 18] を用いる．ReCo は，InfoNCE 損失とは負例に対する最適化の面で異なり，以下に述べる理由から本タスクで適している．

まず，本タスクに InfoNCE 損失を単純に適用した場合，$\boldsymbol{x}_{\text{inst}}$ に対し，1 つの正例 $x_+ = x_t^{(k)}(k = 1, ..., N_{\text{targ}})$ 間の類似度を最大化し，負例の集合 $\mathcal{N} = \{\boldsymbol{x}_-|\boldsymbol{x}_- = \boldsymbol{x}_t^{(m)}, m \neq k, 1 \leq m \leq N_{\text{targ}}\}$ の各要素との類似度を最小化するようにモデルは最適化される．ここで，本タスクにおいて，環境画像群 $\mathcal{C}$ は複数の視点位置から撮影された冗長な画像群であるとみなせる．したがって，$\mathcal{N}$ には異なる視点から撮影された、$\boldsymbol{x}_+$ と同一の対象

物体が含まれる可能性がある．ここで，$s(\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}'_{-}) > 0$ を満たす $\boldsymbol{x}'_{-}$ の集合を $\mathcal{N}'$ とする．すなわち，InfoNCE 損失では，$\mathcal{N}$ と $\mathcal{N}'$ を均等に扱い，$\mathcal{N}'$ の各要素と $\boldsymbol{x}_{\text{inst}}$ の類似度を-1 に近づけるように学習される．この場合，実際には $\boldsymbol{x}_{\text{inst}}$ と類似度が高い物体を強制的に遠ざけてしまうため不適切である．

一方で ReCo 損失は $s(\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}'_{-})$ を-1 ではなく 0 に近づけ，記の問題を軽減する．具体的には，バッチごとに，$S_{ij} = s(\boldsymbol{x}_{\text{inst}}^{(i)}, \boldsymbol{x}_{t}^{(j)})(i, j = 1, ..., |B|)$ を満たす類似度行列を作成する．ここで，$|B|$ はバッチサイズであり，$\boldsymbol{x}_{\text{inst}}^{(i)}$，$\boldsymbol{x}_{t}^{(j)}$ はバッチ内の i，j 番目のサンプルに対応する．S の対角成分と非対角成分を分けて考慮し，それぞれを各バッチにおける正例，負例として扱う．すなわち，以下の形で記述される．

$$\mathcal{L}_{\text{ReCo}} = \sum_{i}(1 - S_{ii})^2 + \lambda \sum_{i} \sum_{j \neq i} \max(0, S_{ij})^2$$

ここで，λ は 2 つの項を補正する正の定数である．本研究の損失は以下で与えられる．

$$\mathcal{L} = \lambda_{\text{InfoNCE}} \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{ReCo}} \mathcal{L}_{\text{ReCo}}$$

ここで，$\mathcal{L}_{\text{InfoNCE}}$，$\lambda_{\text{InfoNCE}}$，$\lambda_{\text{ReCo}}$ は InfoNCE 損失，InfoNCE 損失の重みおよび ReCo 損失の重みを表す．

4. 実験

4.1 データセット

本実験では，データセットとして新たに収集した YAGAMI dataset と，既存のデータセットである LTRRIE dataset [Kaneda 24] を拡張した LTRRIE-subset を扱う．

4.1.1 YAGAMI dataset

公共性の高い建物で撮影された画像群を検索対象とする LTRPO タスクのデータセットは限られている．LTRPO タスクを目的とする既存のデータセットの一つに LTRRIE dataset [Kaneda 24] が挙げられるが，家庭内の生活空間を検索対象としており，また公共空間への応用にも制限がある．具体的には，実用上は人間またはロボットが任意の地点で収集した画像を入力として検索することが理想的であるのにも関わらず，LTRRIE dataset の画像群は，撮影された位置および高さが予め決まっている．このような設定は現実的ではない．

したがって，現実的なデータセットにおいても評価するために，新たに YAGAMI dataset を作成した．YAGAMI dataset は環境画像，対象物体領域，指示文で構成される．現実的な問題設定のもとで実験するために，大学のキャンパスを巡回し，モバイルカメラを用いて環境画像を 10041 枚撮影した．撮影した環境画像に対し，Detic [Zhou 22] によって検出された矩形領域の一部を対象物体領域とした．

また，Detic によって付与した対象物体領域の中には，物体全体を正しく囲えていない，物体が小さすぎて判別ができない，あるいは物体の一部が画角外になるサンプルが存在したため，以下を満たすものを除外した．

- 確信度が 50%未満のサンプル
- 矩形領域の面積が 6000 ピクセル以下のサンプル
- 画像の縁から 3%以内の領域に矩形領域のいずれかの頂点が含まれるサンプル

指示文の収集には，クラウドソーシングを活用した．アノテータにはランダムに並び替えた対象物体領域と環境画像を表示し，物体に作用する指示文を付与するよう指示した．

YAGAMI dataset は，11 の環境，1990 の指示文および 1984 の対象物体領域から構成される．指示文の語彙サイズは 1843，全単語数は 26205，平均文長は 13.17 である．これらの指示文は，57 名のアノテータによって収集された．

4.1.2 LTRRIE-subset

また，多様な環境からデータを収集するため，LTRRIE dataset を拡張した LTRRIE-subset も用いて評価を行った．既存のデータセットは，環境単位で物体の検索を行っており，本研究で扱う大規模環境には適さない．大規模環境を擬似的に再現するために，複数の環境から検索できるように拡張した．LTRRIE-subset には，67 の環境，4930 の指示文および 4709 の対象物体領域が含まれている．指示文の語彙サイズは 2698，全単語数は 92955，平均文長は 18.85 である．

4.2 定量的結果

本研究では，ベースライン手法として CLIP [Radford 21] および MultiRankIt [Kaneda 24] を選択した．CLIP は事前学習で得た知識を fine-tuning を必要とせず text-image retrieval タスクに適用可能であり，優れた結果が得られている．また，MultiRankIt は LTRPO タスクにおいて良好な結果が得られている．以上から，これらの手法をベースライン手法とした．

表 1 に定量的結果を示す．この表では提案手法およびベースライン手法の YAGAMI dataset および LTRRIE-subset それぞれ 2 つのテスト集合に対する性能を比較している．表中の値は 5 回の実験による平均と標準偏差を示している．

表 1 より，YAGAMI dataset において，提案手法は主要尺度である MRR において 25.7%であり，ベースライン手法における最良のスコアを 6.7 ポイント上回った．さらに，提案手法は R@5，R@10 においてそれぞれ 24.5%，36.2%であり，ベースライン手法における最良のスコアを 7.2，5.8 ポイント上回った．また，LTRRIE-subset において，提案手法は主要尺度である MRR において 37.4%であり，ベースライン手法における最良のスコアを 5.0 ポイント上回った．同様に，提案手法は R@5，R@10 においてそれぞれ 35.3%，50.3%であり，ベースライン手法における最良のスコアを 3.4，5.4 ポイント上回った．また，両データセットにおいて，全ての評価指標で有意差があった（$p < 0.05$）．

4.3 Ablation Study

Ablation study として，以下の 3 つの条件を定めた．

ReCo ablation． ReCo 損失の有用性を検証するために，損失関数に InfoNCE 損失のみを使用した．表 2 から，モデル (i) とモデル (ii) を比較すると，モデル (ii) の下では，YAGAMI，LTRRIE のそれぞれで MRR が 0.8 ポイント，0.6 ポイント減少した．この結果から，本タスク設定において，環境中の同一物体を考慮した ReCo 損失が適切なモデルの最適化を可能にしていると考えられる．

Pixel feature ablation． $\boldsymbol{h}_{\text{px}}$ を取り除くことで，性能にどの程度の差が生じるかを調査した．表 2 から，モデル (i) と (iii) を比較すると，モデル (iii) の下では，YAGAMI，LTRRIE のそれぞれで MRR が 2.7 ポイント，5.3 ポイント減少した．これらの結果から，物体とその周辺を含む画素単位での画像特徴が，効果的な接地を可能にしていると考えられる．

Spatial feature ablation． $\boldsymbol{h}_{\text{sp}}$ を除外して，その有効性を検証した．同様に，表 2 から，モデル (i) と (iv) を比較すると，モデル (iv) の下では，YAGAMI，LTRRIE のそれぞれで MRR が 3.4 ポイント，2.4 ポイント減少した．この結果から，中間特徴量の 2 次元の特徴マップから空間的な画像特徴を抽出することで，画像内の物体間の相互関係を理解する能力が向上したことが示唆される．

4.4 定性的結果

図 3 に提案手法の成功例を示す．提案手法および MultiRankIt において，$\boldsymbol{x}_{\text{inst}}$ に対する正解および上位 3 件の検索

表 1: 提案手法とベースライン手法の YAGAMI dataset および LTRRIE-subset における定量的結果

Methods	YAGAMI dataset			LTRRIE-subset		
	MRR [%] ↑	R@5 [%] ↑	R@10 [%] ↑	MRR [%] ↑	R@5 [%] ↑	R@10 [%] ↑
CLIP [Radford 21]	19.0	16.4	25.4	31.6	29.8	42.4
MultiRankIt [Kaneda 24]	18.7±0.8	17.3±1.3	30.4±1.1	32.4±2.9	31.9±1.8	44.9±2.6
Ours	**25.7±1.7**	**24.5±1.1**	**36.2±1.7**	**37.4±2.2**	**35.3±2.1**	**50.3±3.8**

表 2: YAGAMI dataset と LTRRIE-subset における Ablation study の結果

Model	Conditions			YAGAMI dataset			LTRRIE-subset		
	ReCo	$\boldsymbol{h}_{\text{px}}$	$\boldsymbol{h}_{\text{sp}}$	MRR [%] ↑	R@5 [%] ↑	R@10 [%] ↑	MRR [%] ↑	R@5 [%] ↑	R@10 [%] ↑
(i)	✓	✓	✓	**25.7±1.7**	**24.5±1.1**	**36.2±1.7**	**37.4±2.2**	**35.3±2.1**	50.3±3.8
(ii)		✓	✓	24.9±1.3	23.6±2.0	35.2±1.4	36.8±2.3	34.8±1.9	51.3±2.4
(iii)	✓		✓	23.0±1.0	22.7±2.7	34.8±2.5	32.1±2.7	29.9±1.9	45.2±1.2
(iv)	✓	✓		22.3±1.1	21.0±1.3	31.5±1.7	35.0±0.9	35.0±1.2	**51.5±1.9**

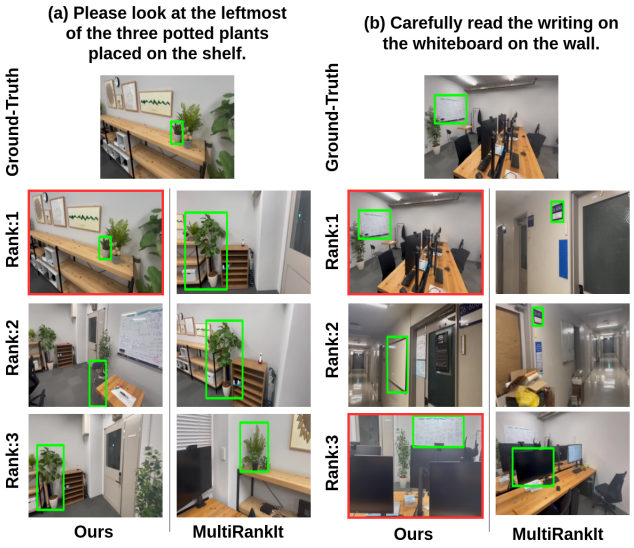

図 3: YAGAMI dataset における定性的結果

結果を表している．緑色で囲まれている領域が $\boldsymbol{x}_t$，赤く囲まれている画像は正解の $\boldsymbol{x}_c$ を表す．

図 3(a) では，$\boldsymbol{x}_{\text{inst}}$ として “Please look at the leftmost of the three potted plants placed on the shelf.” を入力した場合の結果を表している．MultiRankIt では 16 位に検索されている対象が，提案手法では 1 位に検索されている．指示文中の，“the leftmost of the three potted plants” という空間的な位置関係を提案手法は正確に把握されており，これは $\boldsymbol{h}_{\text{sp}}$ が効果的であったと考えられる．さらに，MultiRankIt の上位 3 件も potted plants を検索しているにも関わらず，正解を正しく検索できていない．一方で，提案手法では正しいサンプルを予測できている．これは，これらのサンプルが，$\mathcal{N}'$ に属しているため， RCS で計算した ReCo 損失を考慮した損失関数が有効であったと考えられる．

同様に，図 3(b) は，$\boldsymbol{x}_{\text{inst}}$ として “Carefully read the writing on the whiteboard on the wall.” を与えた場合の結果を表している．MultiRankIt では 6 位と 8 位に検索されているのに対し，提案手法ではそれぞれ 1 位と 3 位に検索されている．MultiRankIt で 1 位と 2 位にはネームプレートが検索されている，この原因として，これらのネームプレートの白い領域を，ホワイトボードと誤って上位に検索されたと考えられる．これを踏まえると，画素単位で物体の意味的な特徴を扱う $\boldsymbol{h}_{\text{px}}$ が効果的であったと考えられる．

5. おわりに

本研究では，大規模な屋内環境内に存在する物体に関する指示文から対象物体を検索し，物体のランク付けリストを出力する LTRPO タスクを扱った．提案手法の貢献は以下の通りである．

- 画像，対象物体，相対位置，画素の 4 つの粒度から，対象物体とその周辺に関する特徴を扱う GREP モジュールを提案した．
- 巡回によって撮影された画像群に含まれる冗長な画像に対しても効率的に学習を可能にする RCS モジュールによる損失計算を導入した．

謝辞

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

参考文献

[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)

[Korekata 23] Korekata, R., et al.: Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks, in *IROS*, pp. 3865–3872 (2023)

[Lin 23] Lin, Z., et al.: Relaxing Contrastiveness in Multimodal Representation Learning, in *WACV*, pp. 2227–2236 (2023)

[Oord 18] Oord, A. v. d., Li, Y., and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018)

[Radford 21] Radford, A., et al.: Learning Transferable Visual Models from Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Schuster 16] Schuster, S., et al.: Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks, in *LREC*, pp. 2371–2378 (2016)

[Vaswani 17] Vaswani, A., et al.: Attention is All You Need, in *NIPS*, pp. 5998–6008 (2017)

[Xu 23] Xu, M., Zhang, Z., Wei, F., Hu, H., and Bai, X.: Side Adapter Network for Open-vocabulary Semantic Segmentation, in *CVPR*, pp. 2945–2954 (2023)

[Zhou 22] Zhou, X., et al.: Detecting Twenty-thousand Classes using Image-level Supervision, in *ECCV*, pp. 350–368 (2022)