マルチモーダル基盤モデルと最適輸送を用いたポリゴンマッチングによる参照表現セグメンテーション

Referring Expression Segmentation using Optimal Transport Polygon Matching with Multimodal Foundation Models

西村 喬行 Takayuki Nishimura　九曜 克之 Katsuyuki Kuyo　神原 元就 Motonari Kambara　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

In modern aging societies, the demand for assistance and support in daily life is increasing; however, there is a feared shortage of home caregivers. Domestic Service Robots, which can be instructed in natural language by care recipients, are gaining significant attention as a solution to improve convenience. In this study, we focus on the task which involves generating segmentation masks of the target object from an image of the indoor environment and the instruction sentence related to object manipulation. Existing methods are unable to handle referring expressions of objects existing outside of the camera's view. Therefore, we introduce Open-Vocabulary 3D Aggregator to obtain open-vocabulary multimodal features about objects existing outside the camera's field of view As a result, our method outperformed the baseline methods on mIoU, P@0.5 and P@0.7 in the OSMI-3D task.

1. はじめに

現代社会で高齢化が進む中，日常生活における介助支援の重要性が高まっているが，その介助を担う在宅介助者は不足している．これの解決策として，被介助者に物理的な支援が可能な生活支援ロボットが注目されている．被介助者が自然言語による指示で生活支援ロボットの物体把持や移動に関する操作ができると便利である．しかし，自然言語による指示文はしばしば複雑な参照表現や冗長な表現を含む場合があり，生活支援ロボットがその様な指示文から対象物体を適切に理解する性能は現状不十分である．本研究では，物体操作に関する命令文が与えられた際，対象物のマスクを生成する Object Segmentation from Manipulation Instructions-3D(OSMI-3D) タスクを扱う．本タスクはロボットの物体把持において重要である．なぜなら，その形状や位置を特定することが重要であり，マスクによる把持物体の領域予測が，矩形領域による予測よりも望ましいためである．例えば，“Go to the living room and bring me the pillow that is closest to the potted plant.” という指示文が与えられた際，potted plant に最も近い pillow のセグメンテーションマスクを生成することが望ましい．

OSMI-3D タスクと関係が深いタスクとして，Referring Expression Segmentation（RES）タスク [Hu 16] がある．本研究で扱う OSMI-3D タスクは，ナビゲーション命令から始まる 2 文以上の指示文が多く，対象物を修飾する句が複数含まれている場合があり，単純な RES タスクよりも困難なタスクである．例えば，指示文 “Go downstairs to the open living room with the white fireplace and straighten out the book display next to it” が与えられたとき，“the book display next to it” のみでは不明瞭であるため，対象物体を特定できない可能性がある．この例では，“the white fireplace” が対象物体を間接的に修飾しているため，その表現の理解が重要になる．OSMI-3D と最も密接に関連しているタスクとして，OSMI タスク [Iioka 23] がある．実際，OSMI タスクでは，OSMI-3D タスクと同様に，与えられた命令は複数の文から構成される．そのため，OSMI モデルの一つである MDSM [Iioka 23] はこのような指示をより適切に理解することができたと報告している．しかし，MDSM はカメラの視界外に存在する物体の参照表現を扱うことができない．また，これらの手法は，一般的にピクセルレベルまたはポリゴンベースのマスクを作成し，ポリゴンベースのマスク生成モデルは良好な結果が報告されている．ポリゴンベースのマスク生成モデルは，対象物体を表現するポリゴンの頂点を予測するものであり，同じ形状であるが頂点の順序が異なる場合を考慮することができないという問題がある．本研究では，画像，3D 点群，複雑な参照表現を含む命令文から，対象物体のセグメンテーションマスクを生成するモデルを提案する．既存手法との主な違いは，画角外に存在する物体の特徴量を扱う Open-Vocabulary 3D Aggregator(OVA) を導入する点である．また，OVA を導入することで，画角外に存在する物体に関する 3 次元点群の open-vocabulary マルチモーダル特徴量を得ることができ，この特徴量及び画角外の物体を用いて対象物体を修飾している参照表現とを対応付けることが期待される．本研究の独自性は以下である．

図 1: OSMI-3D タスクのシーン例

- 画角外に存在する物体に関する 3 次元点群の open-vocabulary マルチモーダル特徴量を得て参照表現と対応づけを扱う，OVA を導入する．
- 参照表現を用いて言語を媒介として構造化した画像の特徴を扱う Visual Context Interpreter(VCI) を導入する．
- セグメンテーション画像に基づきモデルの物体に対する

連絡先: 西村喬行，慶應義塾大学，
神奈川県横浜市港北区日吉 3-14-1，t-nishimura@keio.jp

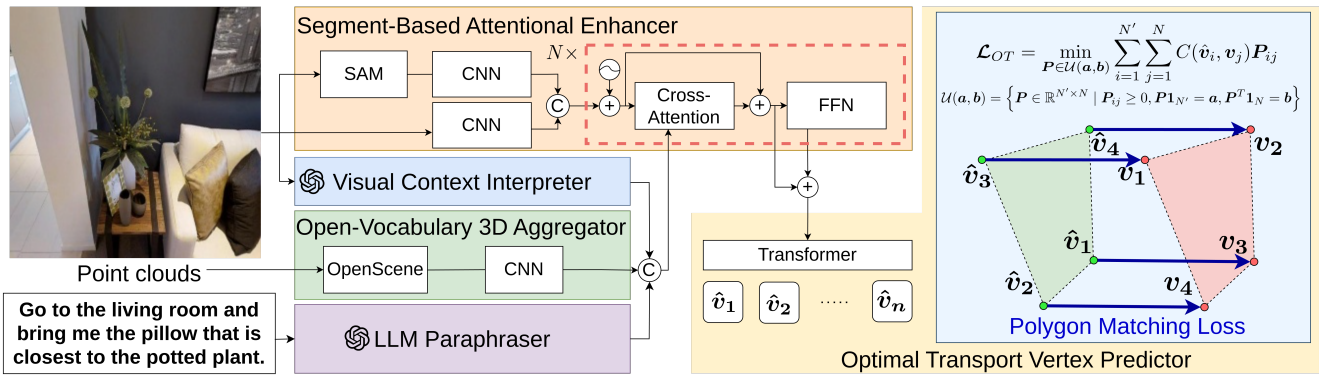

図 2: 提案手法のネットワーク構造

セグメント情報の理解を補助するための Segment Based Attentional Enhancer(SBAE) を導入する．

2. 問題設定

本研究は，屋内環境における画像，3 次元点群及び物体操作に関する指示文から対象物体のセグメンテーションを行うタスクを扱う．ここで本タスクを OSMI-3D タスクと定義する．本タスクでは，与えられた指示文が指す対象物に対して，セグメンテーションマスクを生成することが望ましい．

図 1 に本タスクの具体例を示す．例えば，“Go to the living room and bring me the pillow that is closest to the potted plant.” という指示文が与えられた際，赤色の領域で示すマスクの生成を目標とする．入力は画像，3 次元点群及び指示文である．出力は指示文で指定された対象物体に対するセグメンテーションマスクである．また，本研究では，画像中に対象物が複数ある場合や全くない場合を扱わない．

3. 提案手法

提案手法は与えられた指示文が指す対象物に対するマスクを予測する OSMI-3D を扱う．本モデルの主要モジュールは，Paraphraser，SBAE，Open-Vocabulary 3D Aggregator，VCI，Optimal Transport Vertex Predictor(OTVP) の 5 つである．入力は $\boldsymbol{x} = \{\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{pcl}}, \boldsymbol{x}_{\text{inst}}\}$ である．ここに，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times H\times W}$，$\boldsymbol{x}_{\text{pcl}} = \{p_i \mid i = 0, 1, 2, \ldots, N_{\text{pcl}}\}$ 及び $\boldsymbol{x}_{\text{inst}} \in \{0, 1\}^{v\times l}$ は，それぞれ画像，3 次元点群及び指示文を表す．同様に，p_i，N_{pcl}，v 及び l は，それぞれ i 番目の点群，点群の総数，指示文の語彙サイズ及び最大トークン数を表す．また，Paraphraser [九曜 24] を用いて，$\boldsymbol{x}_{\text{inst}}$ から対象物体に対する修飾句が明瞭となるような要約文 $\boldsymbol{L}_{\text{p-inst}}$ を生成する．

3.1 Visual Context Interpreter

既存の RES モデルでは，画像特徴を抽出するための方法が主に 2 つある．1 つは，ResNet [He 15] や ViT [Dosovitskiy 20] のような画像エンコーダを使用して，テクスチャやエッジのような視覚特徴を抽出する方法．次に，CLIP [Radford 21]，BLIP [Li 22] などのマルチモーダル画像エンコーダにより，自然言語と整合性のあるマルチモーダル画像特徴を抽出する方法である．しかし，これらの特徴は，複雑な参照表現や空間関係に関連する視覚表現が不足していることがある．

以上より，本手法ではそのような複雑な視覚表現を扱うための VCI を導入する．VCI では，大規模視覚言語モデルを用いて画像に含まれる物体の属性・複雑な修飾関係及び空間関係を記述する．また，VCI を用いることで，大規模視覚言語モデルがもつ常識的な知識を用いて，環境や物体に対する，画像単体からは直接読み取ることのできない付加的な情報を得ることが期待される．例えば，ドアの隙間から外の景色が見えることから，そのドアは玄関のドアであるといった情報が得られると考えられる．加えて，VCI は，プロンプトを変化させることで任意の要素に焦点を当て抽出が可能な，条件付き特徴量抽出器であると言える．

VCI の入力は，$\boldsymbol{x}_{\text{img}}$ 及び $\boldsymbol{L}_{\text{p-inst}}$ であり，出力は文埋め込み $\boldsymbol{h}_{\text{nar}}$ である．まず，大規模視覚言語モデルである gpt-4-vision-preview [Achiam 23] を用いてキャプション $\boldsymbol{L}_{\text{nar}} \in \{0, 1\}^{v_{\text{nar}}\times l_{\text{nar}}}$ を取得する．ここで，v_{nar} 及び l_{nar} は，$\boldsymbol{L}_{\text{nar}}$ の語彙サイズ及び最大トークン数を表す．次に，$\boldsymbol{L}_{\text{nar}}$ 及び $\boldsymbol{L}_{\text{p-inst}}$ にそれぞれ text-embedding-ada-002 [ada] を用いて文埋め込み $\boldsymbol{h}_{\text{nar}} \in \mathbb{R}^{d_{\text{nar}}}$ 及び $\boldsymbol{h}_{\text{p-inst}} \in \mathbb{R}^{d_{\text{p-inst}}}$ を得る．ここで d_{nar} 及び $d_{\text{p-inst}}$ はそれぞれ $\boldsymbol{h}_{\text{nar}}$ 及び $\boldsymbol{h}_{\text{p-inst}}$ の次元数を表す．本研究では，モデルの参照表現理解を補助するため，対象物体に関連する色，形状及び他の物体との相対位置に焦点を当てプロンプトを考案した．

3.2 Open-Vocabulary 3D Aggregator

既存手法では，画角外に存在する物体に関する情報を得ることができないため，画角外の物体に関する参照表現が与えられた場合適切に対象物体を予測することができない．そこで，画角外の物体に関する参照表現の理解を補助するために，Open-Vocabulary 3D Aggregator を導入する．本モジュールは，3 次元点群に open-vocabulary のマルチモーダル特徴量を与え，参照表現と対応づけ，別角度から画像を得ることなく画角外に存在する物体に関する情報を得ることが期待される．

本モジュールにおける入力は $\boldsymbol{x}_{\text{pcl}}$ であり，出力は中間特徴量 $\boldsymbol{h}_{\text{pcl}} \in \mathbb{R}^{d_{\text{pcl}}}$ である．ここで，d_{pcl} は $\boldsymbol{h}_{\text{pcl}}$ の次元数を示す．

まず，$\boldsymbol{x}_{\text{pcl}}$ のうち，$\boldsymbol{x}_{\text{img}}$ を取得した位置に最も近い N_{near} 点を抽出し，$\mathcal{P}_{\text{near}}$ とする．ここで，N_{near} 点のみを利用するのは，参照表現は対象物体の周囲の物体に関連する場合が多く，離れた位置にある点を利用することは効率的でないためである．次に，OpenScene [Peng 23] の学習済みモデルを用いて，$\mathcal{P}_{\text{near}}$ から特徴量を抽出する．OpenScene は，3 次元点群の各点に対して CLIP [Radford 21] を用いたマルチモーダル特徴量を埋め込んでいる．最終的に，アップサンプリング及び最大プーリングを行い特徴量 $\boldsymbol{h}_{\text{pcl}}$ を得る．

3.3 Segment Based Attentional Enhancer

既存の OSMI タスクを扱う手法は，物体の輪郭を誤って予測する場合がある．そこで，物体に対するセグメント情報の理解を補助するために，SBAE を導入する．本モジュールは，SAM [Kirillov 23] を用いて生成したセグメンテーション画像から複数の解像度の画像特徴量を抽出し，さらに各モジュールから出力された，特徴量群と統合することでマルチモーダル特徴量を出力する．本モジュールへの入力は $\boldsymbol{x}_{\text{img}}$，$\boldsymbol{h}_{\text{p-inst}}$

表 1: ベースライン手法との比較および Ablation Study の定量的結果

Model	VCI	OVA	SBAE	mIoU	P@0.5	P@0.7
LAVT [Yang 22]				28.16 ± 2.85	26.46 ± 4.01	18.75 ± 3.29
SeqTR [Zhu 22]				21.84 ± 2.24	17.87 ± 7.00	5.16 ± 5.26
MDSM [Iioka 23]				24.36 ± 3.87	22.49 ± 5.46	13.71 ± 3.34
Ours (i)		✓	✓	35.27 ± 5.41	45.31 ± 19.48	19.48 ± 4.99
Ours (ii)	✓		✓	37.36 ± 2.55	48.11 ± 4.13	27.24 ± 4.99
Ours (iii)	✓	✓		31.77 ± 0.92	37.86 ± 2.06	14.00 ± 4.28
Ours (iv)	✓	✓	✓	**38.16 ± 2.46**	**48.85 ± 2.70**	**22.29 ± 3.32**

, $\boldsymbol{h}_{\text{nar}}$，及び $\boldsymbol{h}_{\text{pcl}}$ であり，出力は $\boldsymbol{S_a}$ である．まず，$\boldsymbol{x}_{\text{img}}$ に対して事前学習済みの SAM を利用してセグメンテーション画像を取得し $\boldsymbol{s}$ とする．次に，$\boldsymbol{x}_{\text{img}}$ から MS-COCO [Lin 14] において事前学習済みの DarkNet-53 [Wang 21] を用いて，解像度の異なる M_v 種類の中間層における画像特徴量 $\{\boldsymbol{V}_i\}_{i=1}^{M_v}$ を得る．同様に $\boldsymbol{s}$ から M_s 種類の中間層における画像特徴量 $\{\boldsymbol{V}_i'\}_{i=1}^{M_s}$ を得る．ここに，$\boldsymbol{V}_i, \boldsymbol{V}_i' \in R^{H_i \times W_i \times C_i}$ であり，H_i，W_i，C_i は $\boldsymbol{V}_i$ の画像特徴量のサイズおよびチャネル数を示す．各 $\boldsymbol{V}_i$ に及び $\boldsymbol{V}_i'$ にダウンサンプリングを行った後，図のようにそれぞれをチャネル方向に結合することで $\boldsymbol{V}_{mix}$ を得る．

次に，$\boldsymbol{h}_{\text{p-inst}}$，$\boldsymbol{h}_{\text{nar}}$，$\boldsymbol{h}_{\text{pcl}}$ をチャネル方向に結合後ダウンサンプリングを行い $\boldsymbol{h}_{\text{mix}}$ を得る．$\boldsymbol{V}_{\text{mix}}$ 及び $\boldsymbol{h}_{\text{mix}}$ に対して cross-attention を適用して，$\boldsymbol{S}_{\text{a}} = \{f_{\text{a}}(\boldsymbol{V}_{\text{mix}}, \boldsymbol{h}_{\text{mix}}) | j = 1, \ldots, A\}$ を算出する．ここで，Attention 機構の Head 数を A とする．また，$f_a(\cdot, \cdot)$ は cross-attention を表し，任意の行列 $\boldsymbol{X}_A$ および $\boldsymbol{X}_B$ に対して次のように定義する．

$$f_a(\boldsymbol{X}_{\text{A}}, \boldsymbol{X}_{\text{B}}) = \text{softmax}\left(\frac{(\boldsymbol{W}_q\boldsymbol{X}_A)(\boldsymbol{W}_k\boldsymbol{X}_B)^\top}{\sqrt{d}}\right)(\boldsymbol{W}_v\boldsymbol{X}_{\text{B}})$$

ここで，$\boldsymbol{W}_q$， $\boldsymbol{W}_k$， $\boldsymbol{W}_v$ は学習可能な重み，d はスケーリングファクタである．最後に，OTVP [九曜 24] を用いる．これにより，頂点数 N' 及び頂点の埋め込み表現 $\boldsymbol{S}_a$ から，予測マスクの頂点集合 $\hat{\boldsymbol{\mathcal{Y}}} = \left\{\hat{\boldsymbol{v}}_i \in \mathbb{R}^2\right\}_{i=1}^{N'}$ を出力する．ここで，$\hat{\boldsymbol{v}}_i$ は多角形を構成する頂点の座標を表す．

4. 実験

4.1 データセット

本研究では，REVERIE データセット [Qi 20] および Matterport3D データセットを基に作成された SHIMRIEv2 [九曜 24] を拡張し，SHIMRIE-3D を構築した．SHIMRIE-3D には，画像，Matterport3D [Chang 18] における 3 次元点群，対象物に関する指示文及び対象物に対する多角形に基づくマスクが含まれる．しかし，SHIMRIEv2 には，OSMI-3D タスクの要素である 3 次元点群が含まれておらず，SHIMRIE-3D では Matterport3D における 3 次元点群を導入した．

4.1.1 SHIMRIE-3D データセット

SHIMRIE-3D は REVERIE データセット [Qi 20] および Matterport3D データセット [Chang 18] を基に収集された．まず，これらの指示文について，REVERIE データセットから収集した．REVERIE データセットにおける指示文は，屋内環境中の離れた位置にある物体に対する操作に関するものである．次に，SHIMRIE-3D における対象物のマスクの作成は，ボクセル単位での物体のクラス情報及び対象物体を囲む矩形領域を用いて手動で行った．ここで，ボクセル単位での物体のクラス情報は，Matterport3D データセットに含まれている．また，対象物体を囲む矩形領域は REVERIE データセットに含まれている．加えて，SHIMRIE-3D における 3 次元点群は，Matterport3D データセットに含まれているものを用いた．我々の知る限り OSMI-3D タスクにおいて求められる情報全てを含むデータセットは存在しない．ここで，OSMI-3D タスクにおけるデータセットの要件は，屋内環境における画像，3 次元点群，対象物のマスク及び家事タスクに関する指示文の全てを含むことである．例えば，REVERIE データセットは，実世界の屋内環境を扱った object localization のための標準データセットであり，本研究と関連が深い．しかし，このデータセットには対象物のマスクが含まれていないため，OSMI-3D タスクで使用するには不十分である．また，SHIMRIE データセットも OSMI タスクにおけるデータセットの一つである．このデータセットには，対象物のマスクが含まれているが，3 次元点群が含まれていないため，同様に OSMI-3D タスクにおいて不十分である．以上の理由より，本論文では既存のデータセットは利用せず，対象物のマスク及び 3 次元点群を含む SHIMRIE-3D を新たに構築した．

前処理として，640 × 480 の元画像を 256 × 256 にリサイズした．SHIMRIE-3D には，4,341 枚の画像，11,371 の指示文及びそれに対応する対象物のマスクが含まれている．指示文の語彙サイズは 3,558，全単語数は 196,541 語，平均文長は 18.8 である．SHIMRIE-3D は，全 11,371 サンプルであり，訓練集合，検証集合，テスト集合のサンプル数はそれぞれ 10,153，856，362 である．分割の方法は REVERIE データセットを踏襲しており，既知の環境と未知の環境に分割された 90 のフロアマップから収集し，検証集合において，既知集合，未知集合のサンプル数はそれぞれ 582，274 であり，テスト集合は全て未知集合で構成される．ここに，訓練集合において既知の環境を既知集合，訓練集合において未知の環境を未知集合とする．

4.2 パラメータ設定

損失関数は 79 エポック目までは予測値と Ground Truth(GT) 間の平均絶対誤差，80 エポック目から 90 エポック目までは Polygon Matching Loss [九曜 24] を用いた．学習は，4 時間程度で完了した．また，1 サンプルあたりの推論に要した時間は 6.1ms 程度であった．各エポック毎に検証集合で mIoU の計算を行い，テスト集合での評価は，検証集合で最も高い mIoU を得たモデルを用いて行った．

4.3 定量的結果

定量的結果を表 1 に示す．なお，実験はそれぞれ 5 回行い，その平均及び標準偏差を示した．また，表 1 中の太字は，各尺度において最も高い数値を表す．ベースライン手法は，MDSM，LAVT 及び SeqTR とした．MDSM は OSMI タスクにおいて，LAVT 及び SeqTR は，OSMI タスクと関連の深い RES タスクにおいて，良好な結果が得られたモデルであるためベースライン手法として選択した．本実験における評価尺度には，mean Intersection over Union (mIoU) 及び Precision@k(P@k) を用いた．mIoU 及び P@k は，OSMI-3D タスクと関連の深い OSMI タスク及び RES タスクにおける標準的な尺度であるため使用した．また，本実験の主要尺度は mIoU とした．

表 1 より，LAVT，SeqTR，MDSM，提案手法の mIoU は

“In the 3rd level bathroom, there is a box of tissues to the left of the basin. Please fetch them here.”

図 3: 各手法における定性的結果. 左列から順番に $\boldsymbol{x}_{\text{img}}$，GT，SeqTR における予測マスク，LAVT における予測マスク，MDSM における予測マスク，提案手法における予測マスク.

それぞれ 28.16%，21.84%，24.36%，38.16%であった. 提案手法は，最良の結果が得られた LAVT と比較して 10.00 ポイント上回った．さらに，LAVT，SeqTR，MDSM，提案手法の P@0.5 は 26.46%，17.87%，22.49%，48.85%であり，提案手法は，P@0.5 において最も高い性能である LAVT と比較して 22.39 ポイント上回った．同様に P@0.7 においても提案手法は，それぞれの手法を上回る性能であり，mIoU 及び P@0.5 における提案手法と各既存手法との性能差は統計有意であった.

4.4 定性的結果

図 3 に提案手法における定性的結果の成功例を示す．図左から $\boldsymbol{x}_{\text{img}}$，正解マスク，SeqTR における予測マスク，LAVT における予測マスク，MDSM における予測マスク，提案手法における予測マスクをそれぞれ示す．図 3 において，指示文は “In the 3rd level bathroom， there is a box of tissues to the left of the basin. Please fetch them here” であった．この例では，LAVT と MDSM はマスクを生成せず，SeqTR は雑誌に対して誤ってマスクを生成したが，提案手法はティッシュボックスにマスクを適切に生成した．これより，指示文に含まれる対象物体を適切に理解できたことが示唆される.

4.5 Ablation Studies

Ablation 条件として以下の 3 つを定めた.

VCI ablation

VCI を取り除き，性能への寄与を調査した．表 1 より，モデル (i) における mIoU は 35.27%であり，モデル (iv) よりも 2.89 ポイント減少した．また，同様に P@0.5 においても減少した．これから，VCI が性能向上に寄与しており，VCI により高次の画像理解を補助したことでモデルの参照表現理解の性能が向上したことが示唆される.

OVA ablation

OVA を取り除き，性能への影響を調査した．表 1 より，モデル (ii) における mIoU は 37.36%であり，モデル (iv) よりも 0.8 ポイント減少し，同様に P@0.5 においても減少した.

SBAE ablation

SBAE 内の SAM モジュールを取り除き，それの有効性を調査した．表 1 より，モデル (iii) における mIoU は 31.77%であり，モデル (iv) よりも 6.39 ポイント減少し，同様に P@0.5, P@0.7 においても減少した．これより，SBAE により物体に対するセグメント情報の理解を補助するために，物体の輪郭をより適切に予測することができたことが示唆される.

5. おわりに

本研究では，OSMI-3D を扱った．室内環境の画像，3D 点群，および対象物の操作に関連する指示文から，対象物のセグメンテーションマスクを生成するモデルを生成するタスクである．提案手法による貢献は以下である.

- 画角外に存在する物体に関する 3 次元点群の Open-vocabulary マルチモーダル特徴量を得て参照表現と対応づけを扱う，OVA を導入する.
- 参照表現を手がかりに言語を媒介として構造化した画像の特徴を扱う VCI を導入する.
- セグメンテーション画像に基づきモデルの物体に対するセグメント情報の理解を補助するための SBAE を導入する.
- 提案手法は，mIoU および P@0.5，P@0.7 においてベースライン手法を上回る結果を得た.

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

[Achiam 23] Achiam, J., Adler, S., Agarwal, S., et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2023)

[ada] `https://platform.openai.com/docs/models/embeddings`

[Chang 18] Chang, A., Dai, A., Funkhouser, T., et al.: Matterport3D: Learning from RGB-D Data in Indoor Environments, in *3DV*, pp. 667–676 (2018)

[Dosovitskiy 20] Dosovitskiy, A., Beyer, L., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in *IEEE ICLR*, pp. 12888–12900 (2020)

[He 15] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *CVPR*, pp. 770–778 (2015)

[Hu 16] Hu, R., Rohrbach, M., et al.: Segmentation from Natural Language Expressions, in *ECCV*, pp. 108–124 (2016)

[Iioka 23] Iioka, Y., Yoshida, Y., et al.: Multimodal Diffusion Segmentation Model for Object Segmentation from Manipulation Instructions , in *IROS*, pp. 7590–7597 (2023)

[Kirillov 23] Kirillov, A., Mintun, E., Ravi, N., Mao, H., et al.: Segment Anything, in *ICCV*, pp. 4015–4026 (2023)

[Li 22] Li, J., Li, D., Xiong, C., and Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, in *ICML* (2022)

[Lin 14] Lin, T., Maire, M., et al.: Microsoft COCO: Common Objects in Context, in *ECCV*, pp. 740–755 (2014)

[Peng 23] Peng, S., et al.: OpenScene: 3D Scene Understanding with Open Vocabularies, in *CVPR*, pp. 815–824 (2023)

[Qi 20] Qi, Y., Wu, Q., Anderson, P., et al.: REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments, in *CVPR*, pp. 9982–9991 (2020)

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

[Wang 21] Wang, C., et al.: Scaled-YOLOv4: Scaling Cross Stage Partial Network, in *CVPR*, pp. 13029–13038 (2021)

[Yang 22] Yang, Z., Wang, J., et al.: LAVT: Language-Aware Vision Transformer for Referring Image Segmentation, in *CVPR*, pp. 18155–18165 (2022)

[Zhu 22] Zhu, C., et al.: SeqTR: A Simple yet Universal Network for Visual Grounding, in *ECCV*, pp. 598–615 (2022)

[九曜 24] 九曜 克之, 飯岡 雄偉, 杉浦 孔明：PORTER: 最適輸送を用いた Polygon Matching に基づく参照表現セグメンテーション, 第 30 回言語処理学会資料 (2024)