ARKitSceneRefer: 3D 屋内シーンでの参照表現による小物の位置特定

加藤 駿弥 [1]　栗田 修平 [2]　Chenhui Chu[1]　黒橋 禎夫 [1]
[1] 京都大学大学院情報学研究科 [2] 理化学研究所 AIP
{s-kato, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp　shuhei.kurita@riken.jp

概要

3D 参照表現理解は 3D シーン上でテキストが表す物体の位置を特定するタスクである．しかし，既存の屋内 3D 参照表現理解データセットは主にサイズが大きく，容易に位置特定できる物体を対象としており，ほとんど小物を扱っていなかった．そこで我々は多様で高解像度な ARKitScenes を基とし，ARKitSceneRefer を構築する．ARKitSceneRefer は 1,615 シーン中の，32,300 の屋内で使用する小物から構成され，それぞれ参照表現がアノテーションされている．さらに 2D と 3D の state-of-the-art モデルを用いて実験を行い，その結果，ARKitSceneRefer は挑戦的なタスク設定であることを報告する．

1　はじめに

3D 参照表現理解は 3D シーンを理解する上で重要なタスクであり，物探しロボットなどへの応用が期待される．代表的な 3D 参照表現理解データセットに ScanRefer [1], ReferIt3D [2] が挙げられる．これらのデータセットは机，椅子，ソファーのような屋内の家具クラスなど，比較的に大きい物体を対象にクラスカテゴリが付与されている．しかし，家事などの応用を想定すると，より小さな物体を扱うことが多いと考えられる．例えば，物探しロボットに机の捜索を依頼する人はいない．また料理ロボットは食材や包丁などを扱うことが必要である．

そこで本研究では，ARKitSceneRefer という小物を主な対象とした新しい 3D 参照表現理解データセットを提案する．ARKitSceneRefer は大規模な屋内 3D データセットである ARKitScenes [3] に基づいている．ARKitScenes は ScanRefer と ReferIt3D の基である ScanNet [4] より高解像度で，多様な 3D シーンデータセットである．ScanRefer と ReferIt3D では ScanNet でアノテーションされている物体のみを対象としている．一方で，ARKitScenes には小物の位置がアノテーションされていないので，先行研究とは違い，以下のように小物の位置特定からデータセットを構築する．まず，動画フレームに対して 2D 物体検出を行うことで，小物を抽出する．次に，抽出された小物を，カメラ視点位置から 3D シーンへのレイキャスティングによる位置予測とクラスタリングを用いて 3D シーン上にマッピングすることで，3D 上での小物の座標とクラスを得る．最後に，このデータに参照表現をアノテーションする．加えて，物体検出器から得られたクラスは誤っている可能性があるので，その場合はアノテーション時に修正する．最終的に，ARKitSceneRefer は 731 クラスの物体を対象にし，1,615 シーンに 32,300 の参照表現から構成される．図 1 に ARKitSceneRefer の例を示す．

図 1　ARKitSceneRefer のデータの例．3D シーンと拡大した動画フレームを示す．

データセット評価のため，2D と 3D の state-of-the-art モデル [5, 6, 1, 7] を用いた．実験の結果，ARKitSceneRefer における小物の位置推定は，2D モデル，3D モデルともに挑戦的であることが示された．また，2D モデルにおいては対象物が写っていない動画フレームに対する検出結果がノイズとなっており，うまく対象物が写っている動画フレームの

図 2　ARKitSceneRefer 構築のパイプライン

みを使用することができれば，精度の向上に繋がることも示された．

2　データセット

2.1　データセット構築

本節では，ARKitSceneRefer の構築について述べる．3D シーンより画像のほうが解像度が高く小物を認識しやすいので，本論文では 3D シーン構築の基となった動画フレームから小物を認識する．また，ARKitSceneRefer は bounding box ではなく物体の中心点を予測するタスクを扱う．

2.1.1　物体検出

まず初めに，各シーンを構築する元となった動画を使い，動画のフレームに対して物体検出を行う．物体検出の結果はデータセットの対象となる小物を選別するために利用する．物体検出モデルには LVIS [8] で事前学習された Detic [9] を用いる．Detic から，MSCOCO [10] などで学習された従来のモデルより詳細なクラス情報を得ることができる．時系列的に近い画像にはほとんど同じ物体が写っているため，全フレームに対して物体検出する必要がない．よって 10 フレームごとに物体検出をする．

2.1.2　2D から 3D へのマッピング

Detic からの bounding box を得た後，中心点を動画フレーム上の座標から 3D シーン上の座標にマッピングする．ここで，カメラと物の距離が未知であることと，複数のフレームから同一物体をマッピングしてもカメラパラメータのノイズにより，必ずしも 1 点に収束しないことが問題となる．そこで，我々はレイキャスティングとクラスタリングによってマッピングする．まず，カメラ視点から 3D シーンに対してレイキャスティングを行い，交点を求める．そして，DBSCAN [11] によってこれらの交点をクラスタリングし，各物体の中心点を表すクラスタを作成する．ここでは，クラスタの距離の最大値を 0.05m，最小構成点数を 3 とした．クラスタリングの結果，検証データセットにおいて平均 68.25 のクラスタが得られる．

2.1.3　物体選択

物体選択はクラス選択, シーン選択, 対象物体選択の 3 つのステップからなる．まず，物体検出の結果には色々なサイズの物が混ざっているので，対象となる小物を選択するため，クラス選択を行う．検出されたすべての物体のクラスを見て，両手で掴める程度の大きさ (例: コーヒーメーカー) を目安として，小物となるクラスを選ぶ．この結果，クラス数は 1151 から 794 となる．次に，シーンの多様性と均等性のため，シーン選択を行う．アノテーションするシーンは同じ部屋ごとに 1 つとし，同じ部屋の中で最も物体が多く，かつ物体が 20 以上あるシーンをアノテーション対象とする．この結果，シーン数は 5,047 から 1,615 となる．最後に，アノテーションする物を決めるため，対象物体選択を行う．対象物体数は 1 シーンにつき，20 とする．検出回数が少ないクラスの物体の数を増やすため，ランダムに対象物体を選ぶのではなく，検出回数の低いクラ

表 1　3D 参照表現理解データセットの比較

データセット	環境	参照表現	平均の長さ	シーン数	シーンごとの物体数	クラス数
Sr3D [2]	ScanNet	83,572	9.68	1,273	6.96	76
Nr3D [2]	ScanNet	41,503	11.32	641	9.17	76
ScanRefer [1]	ScanNet	51,583	20.27	800	13.81	279
Ours	**ARKitScenes**	32,300	10.01	**1,615**	**20**	**731**

スの物体から順に 1 つずつ選ぶ．この結果，データセット全体のクラス数は 612 となる．

2.1.4　アノテーション

上記のプロセスを実行することで，何もアノテーションされてない 3D シーンから，小物の位置情報を手に入れることができる．データセット構築の最後のプロセスとして，Amazon Mechanical Turk[1)] 上でアノテーションを行う．ワーカーは，3D シーン，対象物体の写っている 3 枚の画像，対象物体のクラスを参照できる．画像は bounding box の面積が大きいものの中からランダムに選ばれており，bounding box を可視化することでワーカーが対象物体の場所を把握しやすくなっている．さらにワーカーが対象物体と他の物体を明確に認識しやすいように，3D シーン上にあるすべての物体や各タスクのすべての対象物体の位置が可視化できるようになっている．実際のインターフェースは付録を参照されたい．ワーカーはこれらの情報を参考にしながら，対象物体に対して，明確にそれらが区別できるように，参照表現をアノテーションする．さらに，Detic から得られたクラスは間違っている場合もあるので，ワーカーは，クラスが間違っている場合は正しいクラスを選ぶようにする．このときの語彙は Detic と同様の LVIS を用いる．クラス修正の結果，ワーカーは 40.72% のクラスを修正し，データセットのクラス数は 731 となった．間違って Detic に小物と認識されたものがワーカーによって小物ではないクラスに修正されたデータが 4.80% あるので，すべてのデータが小物ではないことに注意しなければならない．

2.2　データセットの統計

我々は 1,615 の 3D シーンに各 20 ずつ，計 32,300 もの参照表現が付与されている小物を対象としたデータセットを構築した．各物体には平均で 10.01 語の参照表現が付与されている．参照表現は屋内の物体の 731 クラスを含んでいる．表 1 に既存の

1)　https://www.mturk.com/

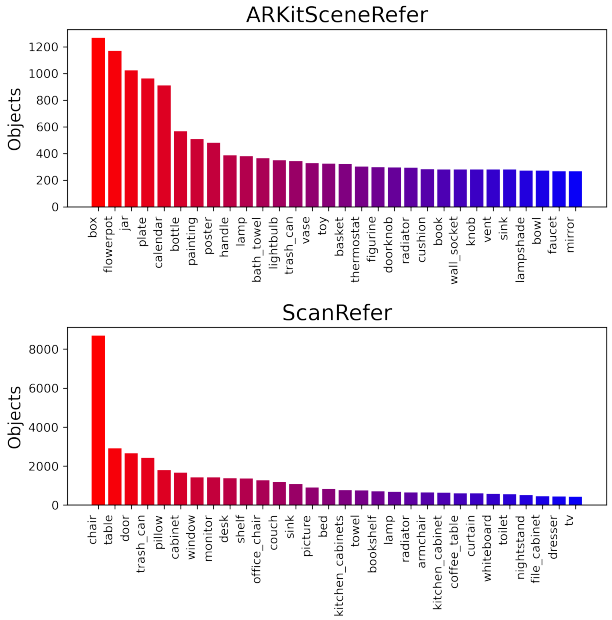

図 3　ARKitSceneRefer (上) と ScanRefer (下) の頻度の高いクラスの分布

3D 参照表現データセットとの比較を示す．我々のデータセットは 3D シーンとクラスの 2 つの点で，既存のデータセットより大規模である．図 3 に ARKitSceneReferr と ScanRefer のクラス分布の比較を示す．例えば，ScanRefer では “table”, “door”など大きい物体のクラスが多くなっている．一方で，ARKitSceneRefer では，“box”, “jar”など ScanRefer と比べると小物のクラスが多いことがわかる．このクラス分布により，我々のデータセットは小物に焦点を置いていることがわかる．

3　実験

3.1　実験設定

本研究では，2D 手法と 3D 手法を用いて評価した．2D 手法は動画フレームと参照表現を入力とし，参照表現に対応する bounding box を予測する．bounding box の中心はレイキャスティングによって 3D シーンにマッピングされ，各動画フレームごとの 3D シーン上の予測座標が得られる．これらをクラスタリングし，点数の最も多いクラスタの中心点

表 2　2D モデルと 3D モデルの比較

手法	距離平均	Acc@0.1	Acc@0.3	Acc@0.5
MDETR-random	2.32	7.77	10.94	13.22
OFA-random	2.26	8.55	11.72	14.05
OFA-Detic	**1.61**	**15.66**	**25.44**	**31.16**
ScanRefer	1.83	1.88	12.44	19.74
3DVG-Transformer	1.68	2.65	16.66	24.89

を最終的な予測とする．一方，3D 手法は点群と参照表現を入力とし，直接参照表現に対応する物体の 3D シーン上の座標を予測する．ARKitSceneRefer では，参照表現に対応する物体の 3D シーン上の中心点を予測するタスクを扱うので，評価指標には点間のユークリッド距離を用いた．Acc@d (%) では距離が d 以下であれば正解とみなした．また，参考として距離平均も評価した．

2D モデルとして，RefCOCOg [12] で fine-tuning された MDETR [5] と OFA [6] を用いた．また以下の 2 つの手法で比較した．

- MDETR-random と OFA-random: 動画フレームをランダムに 1/10 だけサンプリングして入力として使用する．
- OFA-Detic: 動画フレームをランダムにサンプリングした後，Detic を用いて物体検出し，検出クラスが参照表現に含まれる動画フレームのみ入力として使用する．

クラスタリングには DBSCAN [11] を用いた．クラスタの最大距離は 0.02m で，最小構成点数は 1 とした．

3D モデルとして，ScanRefer [1] と 3DVG-Transformer [7] を用いた．点群の頂点数は 200,000 とした．言語特徴量には GloVE [13] を用いた．

3.2　定量的評価

表 2 に ARKitSceneRefer で 2D 手法と 3D 手法を評価した結果を示す．OFA-Detic を除いた 2D 手法と 3D 手法を比較すると，Acc@0.1 においては MDETR-random が 7.77%, OFA-random が 8.55%であるのに対し，ScanRefer が 1.88%，3DVG-Transformer が 2.65% と 2D 手法のほうが優れていたが，その他の評価指標については，例えば Acc@0.5 においては MDETR-random が 13.22%, OFA-random が 14.05% であるのに対し，ScanRefer が 19.74%，3DVG-Transformer が 24.89% であり，3D 手法のほうが優れていた．これは 3D 手法は 3D シーン全体を参照できるので，対象物体が認識できなくてもおおよその位置を当てることができるが，2D 手法は動画フレームに対象物体が写っていない予測がノイズになるからである．また OFA-Detic はその他の手法をかなり上回っていることから，対象物体が写っていると推測される動画フレームのみ入力に使用するとノイズが削減され，精度が向上することが示された．

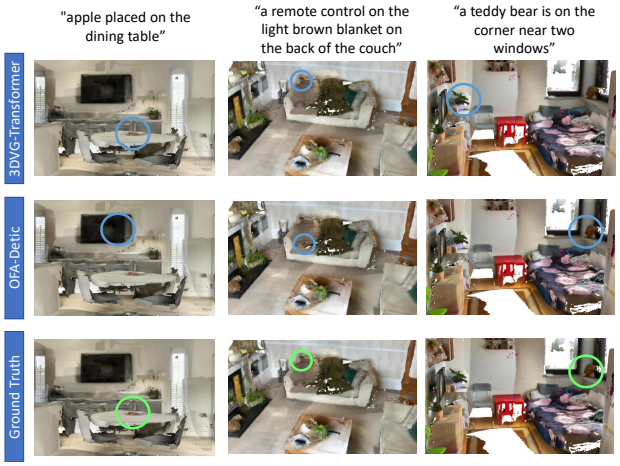

図 4　モデルの出力例

3.3　定性的評価

図 4 に 3DVG-Transformer, OFA-Detic, ground truth の比較結果を示す．左の 2 つの結果は 3DVG-Transformer がグランディングに成功し，OFA-Detic が失敗している例である．前述したように，OFA-Detic の予測精度は入力動画フレームにかなり依存しており，ノイズが多いと無関係な場所を予測してしまう．右の結果は OFA が位置特定に成功し，3DVG-Transformer が失敗している例である．3DVG-Transformer は “on the corner” は捉えられているものの “near two windows”は失敗している．このように，3D 手法は薄いもの (例: タオル) を捉えることができないことがある．一方で画像であれば，正面から撮ることで物の厚みを考慮せず認識できる．

4　おわりに

本論文では，小物の位置特定のための新しい 3D 参照表現理解データセットである，ARKitSceneRefer を紹介した．ARKitSceneRefer には，1,615 の ARKitScenes の 3D シーンに対して 32,300 の参照表現が付与されている．ARKitSceneRefer では，3D モデルはあまり精度が良くなかったが，2D モデルは入力の動画フレーム次第では精度が向上することを確認した．今後の課題として，3D モデルの精度を改善することが挙げられる．

謝辞

本研究は JST さきがけ JPMJPR20C2，JSPS 科研費 22K17983 および サムスン SDS 株式会社の支援を受けたものである．

参考文献

[1] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In **Proceedings of the European Conference on Computer Vision (ECCV)**. Springer, 2020.

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In **Proceedings of the 16th European Conference on Computer Vision (ECCV)**, 2020.

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)**, 2021.

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, 2017.

[5] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 1780–1790, October 2021.

[6] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, **Proceedings of the 39th International Conference on Machine Learning**, Vol. 162 of **Proceedings of Machine Learning Research**, pp. 23318–23340. PMLR, 17–23 Jul 2022.

[7] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 2928–2937, 2021.

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2019.

[9] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In **ECCV**, 2022.

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. **CoRR**, Vol. abs/1504.00325, , 2015.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**, KDD'96, p. 226–231. AAAI Press, 1996.

[12] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In **CVPR**, 2016.

[13] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In **Proceeding of Empirical Methods in Natural Language Processing (EMNLP)**, 2014.

表 3　OFA-oracle と OFA-upper を加えた精度の比較

手法	距離平均	Acc@0.1	Acc@0.3	Acc@0.5
OFA-random	2.26	8.55	11.72	14.05
OFA-Detic	1.61	15.66	25.44	31.16
OFA-oracle	1.51	18.88	30.72	37.33
OFA-upper	**0.37**	**43.83**	**63.22**	**74.72**

図 5　水色の bounding box は OFA のグランディング予測結果を表す．左の例は対象物体を含むが，右の例は含まない．

A　2D 手法の精度の最大値

入力動画フレームによってどのように精度が変化するのか検証するため，新たに以下の 2 つの設定を加えて実験した．

1. OFA-oracle: 動画フレームをランダムにサンプリングした後，Detic を用いて物体検出し，検出クラスがアノテーションされているクラスと一致するなら入力として使用する．
2. OFA-upper: アノテーション時にワーカーに見せた 3 枚の動画フレームのみ入力として使用する．

表 3 はその実験結果を示している．OFA-oracle は OFA-random と OFA-Detic を上回っている．さらに，OFA-upper は OFA-oracle を大きく上回っている．この結果から，動画フレームに対象物体が写っていないと予測がノイズとなり，精度の低下を招くことがわかる．

B　より詳細な 2D モデルの定性的評価

動画フレームに対象物体が含まれていない場合，OFA がどのような出力をするのか検証する．図 5 に対象物体を含む結果と含まない結果を示す．対象物体を含むフレームに対しては正確に予測しているが，対象物体を含まないフレームに対しては関係のない場所を予測している．この誤りがノイズとなり，正確な予測を妨げる．

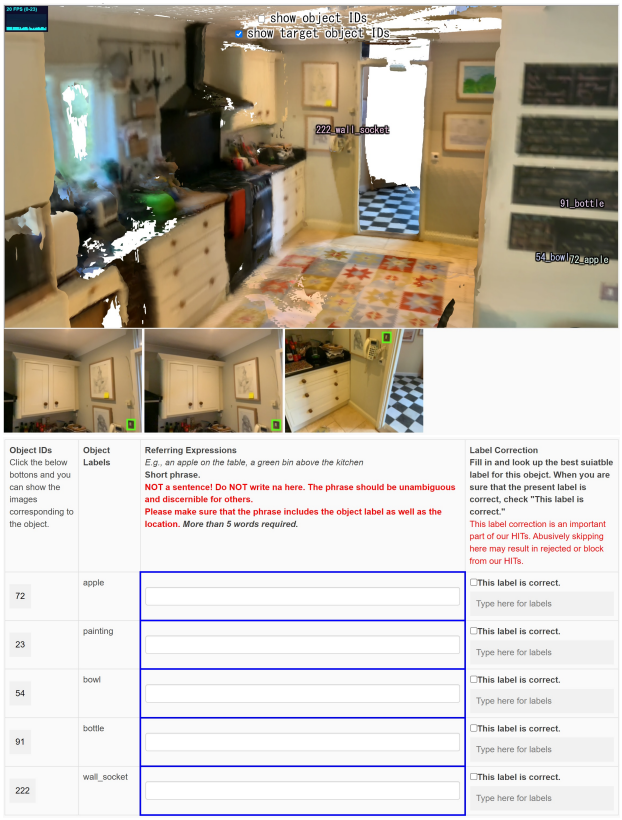

図 6　アノテーションインターフェース

表 4　ARKitSceneRefer の統計

分割	参照表現	シーン数	クラス数
訓練	28,720	1,436	712
検証	1,780	89	328
テスト	1,800	90	349

C　アノテーション

図 6 にアノテーションインターフェースを示す．1 タスクが 5 つの対象物体を含むように設計し，ワーカーの国をアメリカ，カナダ，イギリス，オーストラリアに限定して，アノテーションを行った．また，ワーカーは 6 単語以上の参照表現を書かなければならないように設定した．

D　データセットの分割

表 4 に示されるように，データセットを訓練, 検証, テストに分割した．