2P-06

単眼画像からの三次元シーン再構成に関する研究

仙田　朋也†　　上田　芳弘‡　　坂本　一磨‡
公立小松大学大学院サステイナブルシステム科学研究科生産システム科学専攻†
公立小松大学生産システム科学部‡

１．　はじめに

近年，深層学習の 3D 分野への利用に多く関心が寄せられている．また，自動運転や VR，AR などの 3D を扱った技術の発展は著しく，普及に向けて多くの研究が行われている．さらに，毎年開催されるコンピュータビジョンに関する学会である CVPR(Conference on Computer Vision and Pattern Reconstruct)でも 3D を扱った研究が多くみられ，ボクセル生成，点群生成，メッシュ生成，3D 物体認識など多岐に渡る．その中でも本研究では，画像を用いた 3 次元物体認識と生成に注目する．画像を用いる際，単眼画像か複眼画像を用いる方法があるが，今回は単眼画像からその画像内のシーン理解と再構成に取り組む．シーンの理解と再構成とは，対象となるシーンのレイアウトすなわち，どこに何があるかを理解し，3 次元空間上で画像内のシーンを再現することである．しかし，単眼画像から 3D シーンを理解することは，人間にとっては容易であるが，コンピュータにとっては未だ困難な課題がある．本研究では，NeRF(Neural Radiance Fields)[1]を用いた 3D 生成手法と 3 次元物体認識の手法を組み合わせることで単眼画像からのシーン再構成に取り組む．

２．　提案手法

深層学習分野における 3D 生成を扱った研究の課題として，点群や 3D メッシュなどの 3D 学習データを大量に用意することが困難であるという点が存在する．このことにより，学習に利用できるデータセットが限られてしまう．よって，様々なシーンを再構成する際 3D モデルを生成可能な物体と不可能な物体が存在し，再構成可能なシーンが制限される．これに対処するため既存研究[2]を元に，生成可能カテゴリにない物体を直方体として仮想空間上に配置することを提案した．図 1 はその処理手順の概要である．対象となる画像からシーン内に存在する物体を検出し，検出された物体それぞれの 3 次元位置を推定して，3D メッシュを生成する．その際，生成可能なカテゴリに存在しない物体は，直方体に置き換える．これにより直方体を用いたシーン再構成を可能にしている．しかしこの手法では一部の物体を直方体で置き換えているため対象のシーンを忠実に再現しているとはいえない．本提案手法ではより忠実にシーンを再構成するために，NeRF[1] と CLIP(Contrastive Language-Image Pretraining)[3]を組み合わせたテキストを用いて認識された物体を 3D で生成する手法を用いる．これによって生成された 3D モデルを 3 次元物体認識の推定結果を用いて 3D 空間上に設置することでシーンの再構成を実現する．図 2 は本提案手法の概要を示したものである．図 2 のとおり本手法では再構成を行ないたいシーンを写した 1 枚の画像を入力として用い，最初に 3 次元物体認識を行ないシーン内に存在する物体カテゴリと各物体の 3 次元位置を推定する．次に物体カテゴリを入力テキストとしてテキストによる 3D 生成を行なう．生成された 3D モデルを最初

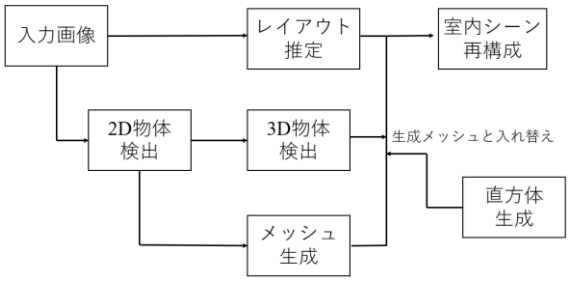


図 1　既存研究概要

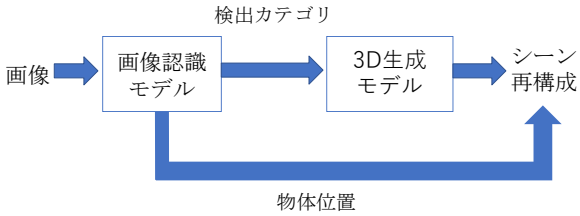


図 2　提案手法概要

Research for 3D Scene Reconstruction from Monocular Camera Image
†Tomoya Senda, Graduate School of Sustainable Systems Science Division of Production System Science, Komatsu University
‡Yoshihiro Ueda and Kazuma Sakamoto, Faculty of Production Systems Engineering and Sciences, Komatsu University

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

に推定したそれぞれの３次元位置情報を元に３次元空間上に配置する．

以上の処理により本手法では３次元のシーン再構成を行なう．図１の既存研究は 3D 生成が不可能な物体を直方体で置き換える処理を加えているが，提案手法においては CLIP を用いた 3D 生成手法を利用することで様々な物体の生成に１つの生成手法のみで対応する．

２.１　3D 生成手法

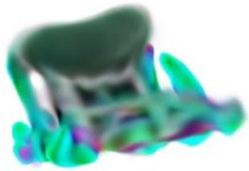
本手法では 3D の学習データを大量に用意することが困難であるという課題に対処するために，3D 生成の手法として NeRF と CLIP を組み合わせたテキストを入力とした 3D 生成[4]を参考にした手法を用いる．NeRF は一つの 3D シーンを生成するために様々な視点からの画像を必要とする．NeRF はその出力結果を用いてレンダリングを行うことで画像を生成することができる．生成された画像と学習データとで損失を取得して学習を行っている．NeRF の学習は 3D データを必要とせず，画像とカメラポーズのみを用いることが特徴である．

一方 CLIP は大規模のデータで学習されたモデルでテキストと画像の類似度を推定することができる．既存研究[4]では NeRF と CLIP の２つを組み合わせることでテキストから 3D モデルの生成を可能としている．NeRF が生成した画像から CLIP を用いて入力テキストとの類似度を推定し，これを損失とする．この損失を元に NeRF の学習を行なうことでテキストからの 3D 生成を可能にしている．CLIP は web 上から取得された大規模なデータセットを用いて学習されており，様々な画像とテキストの類似度を推定することができる．よって、２つの手法の組み合わせによりカテゴリの制限なく様々な物体を 3D 生成することが可能となっている．

３．3D 生成例

[4]の手法を用いて実際に 3D 生成を行った．椅子を生成するためにテキスト「chair」を入力しその結果を確認した．図３の A はその出力結果で B は一般的な椅子の画像例になる．出力結果を一般的な椅子と比較すると，形状が歪んでおり周囲に余計な部分が存在することがわかる．さらに画像中央から上部にかけて背もたれのようなものは確認できるが座面から足にかけての部分がつぶれているように見える．生成結果は入力テキストに対してその形状をうまく表現できていないと考えた．

生成がうまくいかない理由として，入力テキストとして物体カテゴリのみを入力しているためであると考える．今回物体カテゴリのみを入力としているため，入力テキストが簡素になってしまった．それを入力として用いると，ある程度椅子らしいものが生成可能になると損失がうまく取れなくなってしまう．これによって学習がうまくいかず最終的な出力結果が A のようになってしまったと考えた．



A:3D 出力結果

B:一般的な椅子

図３　3D 生成出力の比較

以上のように NeRF と CLIP を使用し，単眼画像から３次元物体を生成した．これを用いることで，存在しない物体を生成可能だと考える．

４．　今後の展望

今回行った 3D 生成では，物体カテゴリのみを入力した．そのため図３の A のような出力結果になったと考えた．よって今後は入力テキストをシーン内に存在する物体を適切に表現するテキストにする必要があると考える．そのために３次元物体認識を行った後に物体カテゴリのテキストに対してシーン内に存在する物体の形状や色に関する情報を付け加える必要があると考えている．

参考文献

[1] Ben M, Pratul P. S, Matthew T, Jonathan T. B, Ravi R, and Ren Ng. : NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, *in European conference on computer vision*, Springer, pp. 405–421, 2020,

[2] Yinyu N，Xiaoguang H，Shihui G，Yujian Z， Jian C， and Jian J. Z. : Total3Dunderstanging: Joint Layout，Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 55-64, 2020.

[3] Alec R, Jong W K, Chris H, Aditya R, Gabriel G, Sandhini A, Girish S, Amanda A, Pamela M, Jack C, Gretchen K, and Ilya S. : Learning transferable visual models from natural language supervision, 2021

[4] Ajay J, Ben M, Jonathan T. B, Pieter A, and Ben P. : Zero-shot text-guided object generation with dream fields, *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 867–876, 2022.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.