2U-09

単眼カメラ画像からの３次元シーン再構築に関する研究

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†　公立小松大学生産システム科学部‡

１．はじめに

単眼カメラ画像から 3 次元の空間を再構築することは，ロボット制御，自動運転，VR や AR などの 3 次元を扱うアプリケーションにおいて基本的なタスクである．画像のみを用いることで簡単に 3 次元を扱うアプリケーションに対して，高価かつ専門的なセンサーを必要とせずに誰でもアクセスが可能となるため非常に有用である．また，3 次元データセットの拡張においても有用であると考える．近年深層学習の分野では大規模なデータセットを用いた大規模学習を行うことによって様々なタスクにおいて大きな成果を上げた．しかし，3 次元を扱った分野では 3 次元データの取得の難しさから，画像やテキストなどの大規模なデータセットを用意することが困難である．単一画像から現実に忠実な 3 次元データを生成することが可能となれば，既存の大規模な画像データセットから 3 次元のデータセットを作成することができ，3 次元を扱うことのできるモデルの大規模学習が可能になると考えている．本研究では単眼カメラ画像のみを入力として画像内に写るシーンを 3 次元データとして再構築することを目指す．

２．提案手法

図 1 が提案手法の概要を図に示したものである．本研究において提案する手法は，3 次元再構築にあたって必要な情報を画像から推定する 1 段階目の処理と，1 段階目において推定した情報を元に 3 次元物体生成を行う 2 段階目の処理，これら 2 つの処理を行いそれぞれの出力結果を用いることで単眼画像からの 3 次元再構築を可能とする．1 段階目と 2 段階目両方において大規模なデータによる事前学習済みのモデルを用いる．1 段階目の処理においては，セグメンテーションモデル

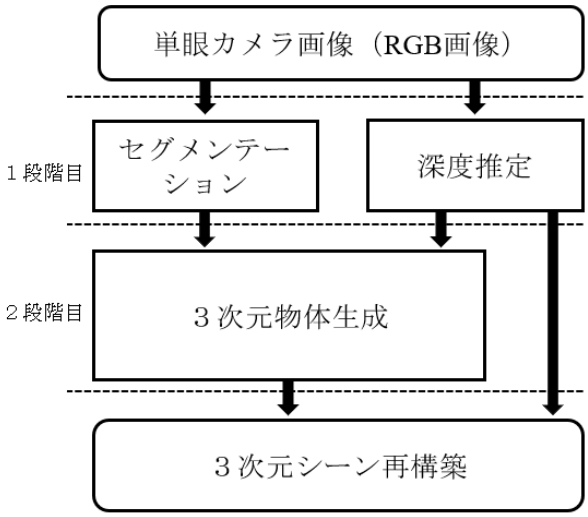

図１　処理概要図

と深度推定モデルの 2 つを用い，2 段階目の処理においては画像生成モデルを事前知識として用いた 3 次元生成手法を採用する．最終的には 1 段階目の処理において得られた物体位置に，2 段階目において生成された 3 次元物体のメッシュデータを空間上に配置することで 3 次元シーン再構築を可能とする．

２．１　1 段階目処理

1 段階目の処理では，シーン内に存在する各物体の画像上におけるピクセル領域とカメラからの距離 2 つの推定を行う．本提案手法においては，画像内より検出された各物体に対して 3 次元物体生成を適応する．今回用いる 3 次元生成は生成する物体の画像が必要であるため，検出された物体の画像上での領域を切り抜く必要がある．そのためにセグメンテーションを用いる．セグメンテーションとは画像上の物体カテゴリ（例：椅子，人，テレビなど）をピクセル単位で推定するタスクである．セグメンテーション結果から，画像内の各物体の領域とそのカテゴリを推定できる．3 次元再構築時に生成した物体を対象シーンに対して忠実に空間上に配置するために，各物体の 3 次元空間内で位置を推定する必要があり，このために深度推定を用いる．深度推定は，カメラからの距離である深度を推定するタスク

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

である．推定された深度画像は入力画像と同じ解像度であるためセグメンテーション結果と1対1で対応しており，この2つのタスクの結果を組み合わせることで深度値から求められた点群データに物体カテゴリと認識番号を結びつけることができ，どの物体がどの位置に存在しているかを推定することができる．

２．２　2段階目処理

2段階目の処理では事前学習済みの画像生成モデルを組み込んだ ImageDream[1]を用いた3次元物体生成を行う．この手法は画像とテキストを用いた3次元生成手法で，入力データから得られる詳細な情報を元に複数の視点から3次元物体の最適化を行う．1段階目のパノプティックセグメンテーション結果の物体ごとの切り抜き画像と検出されたカテゴリを入力とすることで本3次元物体生成を行う．ここで生成した3次元物体を深度推定値の値を元に3次元空間に配置することで3次元シーン再構築を行う．生成された3次元物体とそれに対応する深度値間で ICP(Iterative Closest Point)を行うことで3次元物体の位置と向きを深度値に合わせて調節し，それぞれを適切に空間に配置することで3次元シーン再構築を可能とする．

３．実験結果

パノプティックセグメンテーションとして[2], 深度推定手法として[3]を用いて，3次元シーン再構築を行った．入力画像サンプルは[4]より引用した単眼カメラ画像であり、キッチンを写した画像である．この画像内には椅子，シンク，電子レンジなどが存在している．図 2 がパノプティックセグメンテーションの結果, 図 3 は深度推定結果である．パノプティックセグメンテーションにおいて検出されたカテゴリと画像上におけるその物体領域をもとに3次元物体生成を行い，深度画像の値をもとに生成物体を3次元空間上に配置することで3次元再構築を行った．その最終出力結果が図 4 である．再構築結果では，シーンの中央に椅子が配置され，右側にはシンク等が配置された．しかし，それぞれ配置した物体サイズが元画像のシーンに対して忠実でないためそれぞれが重なりあっていることがわかる．

４．課題と今後の展望

本提案手法では，多種多様な3次元シーンの再構築を目指した．3次元再構築結果である図 4 より，本手法は物体生成と物体配置両方において課題があると考える．特にカメラに対して手前の物体，画像奥の壁や天井といった薄く広い背景となる部分についてはそれぞれ生成手法に改善が必要である．さらに本手法では生成した各物体のスケールを考慮できていないため，元画像と比べるとシーン内に存在する物体のサイズ比を合わせることができない．以上の点に取り組むことでより入力画像に忠実な3次元シーンを再構築可能であると考える．

図2　パノプティックセグメンテーション

図3　深度推定結果

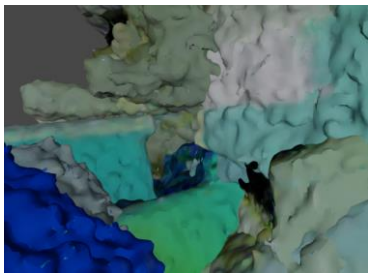

図4　3次元シーン再構築結果

参考文献

[1] Wang, P., and Shi Y. :ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation. *arXiv preprint arXiv:2312.02201*, (2023).

[2] Li, F., Zhang, H., Xu, H., Liu, S., Lei, Z., Ni, L., and Shum, H. : Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. *CVPR*, pp. 3041-3050, (2023).

[3] Oquad, M., Darcet, T., Moutakanni, T., Huy, Vo., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa F., Alaaeldin, E., et al. : Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, (2023).

[4] Song, S., Lichtenberg, S., and Xion, J. :SUN RGB-D: A RGB-D scene understanding benchmark suite. *CVPR 2015*, pp. 567-576, (2015).