NeRF を用いたデータ拡張による APR の精度向上の検討

山村大斗[1]　植村匠[1]　尾島修一[1]

概要：撮像され画面に表示された現実の物体と仮想物体を重畳表示する AR 技術において，単眼カメラのみでカメラポーズ推定を行う手法がある．近年，深層学習でカメラポーズを推定する APR（Absolute Pose Regression）が提案されているが，高い精度を実現するためには大量の多視点画像群が必要であり，それらを手動で得ることは非現実的である．本研究では，学習データを増やすために自由視点画像生成技術を用いることを提案する．今回は，NeRF を用いて生成した新しい視点の画像でデータ拡張を行い，APR の性能向上を目指す．モデルの性能を比較する実験を行った結果，提案手法で学習したモデルが高い性能を示した．

キーワード：仮想／人工／拡張現実，画像認識・理解，機械学習

Improving accuracy of APR through data augmentation using NeRF

DAITO YAMAMURA[†1]　TAKUMI UEMURA[†1]
SHUICHI OJIMA[†1]

Abstract: In augmented reality (AR) systems that capture real-world objects with a camera and overlay virtual objects onto the display, there exist methods that estimate the camera pose using only a monocular camera. Recently, deep learning–based Absolute Pose Regression (APR) has been proposed for camera pose estimation; however, achieving high accuracy requires a large set of multi-view images, and manually collecting such data is impractical. This study proposes to use free-viewpoint image generation techniques to increase the amount of training data. In particular, we perform data augmentation using novel-view images generated with Neural Radiance Fields (NeRF) to improve the performance of APR. Experimental results comparing model performance show that the model trained with the proposed method exhibits higher accuracy.

Keywords: Virtual reality/artificial reality/augmented reality, Image recognition/understanding, Machine learning

1. はじめに

AR（Augmented Reality）とは，現実世界から得られる情報に，仮想世界の情報を加えることで，より豊かな視覚情報を視聴者に与える技術である[1]．その中でも，画面内の現実の物体と，その内部構造等を表す 3DCG モデルとを重ね合わせて表示することは，見えないものを見えるようにするもので，AR を構成する基盤技術の中でも中核となるものである（以下，AR 重畳表示技術と呼ぶ）．この技術は，医療をはじめとした様々な分野で活用されている．図 1 に示すのは，代表的な応用例である外科手術の支援の様子である[2]．画面内の患者の頭部に対して，患部を強調表示する 3DCG を重畳表示するで，医師からみて患部が識別しやすくなり，外科手術の安全性向上に役立つことが期待されている．

AR 重畳表示技術においては，シーンや物体と，それを撮影するカメラの相対的な位置・姿勢（カメラポーズ）を推定する必要がある．このカメラポーズを推定する方式はロケーションベース型とビジョンベース型に大別される．

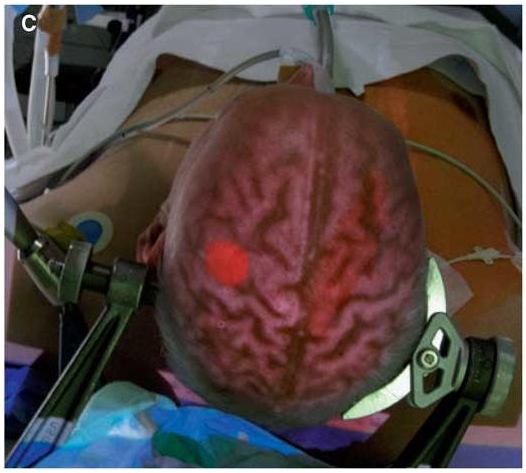

図 1　外科手術における AR 重畳表示技術の活用例 ([2]より引用)

ロケーションベース型では使用するデバイスに搭載された GNSS センサ等から取得した位置情報を利用すること

1 崇城大学
Sojo University

でカメラポーズ推定を行う．一方，ビジョンベース型では，カメラを使ってリアルタイムに撮影された画像中の特徴量などを利用することでカメラポーズ推定を行う．これら 2 つのタイプのうちビジョンベース型は，さらにマーカ型とマーカレス型に分けることができる．

マーカ型では，専用のマーカを撮影対象やその周辺に設置し，それをカメラで撮影することによってマーカの特徴点が検出され，カメラポーズ推定が行われる．この方法では専用に設計されたマーカを用いることによりカメラポーズ推定の精度を上げやすいという利点がある一方で，マーカを用意したり事前に設置したりする手間があることや，カメラの撮影範囲がマーカに制限されてしまうことなどの欠点がある．

このようなマーカ型 AR に対してマーカレス型では，専用マーカに頼らずに，カメラポーズ推定を行う．この場合は，マーカ型にあったマーカ設置の手間や撮影の制約はないため，利用者はより手軽かつ自由に撮影を行うことができる．しかしこの方式の欠点として，マーカ型と比較してカメラポーズ推定の精度を上げることが困難であることが挙げられる．推定精度を上げるために複雑なアルゴリズムを使うと，計算コストが増加して AR としてのリアルタイム性が損なわれることになる．これを解決するために，通常の RGB カメラに加えて 3D 深度センサなどの外部機器を使うことによってカメラポーズ推定の補助に用いる手法がある．そうすることによってより正確なカメラポーズ推定が可能になるが，追加の機器の導入に高いコストがかかったり，準備や撮影の手間が増えたりする欠点がある．以上の背景から，外部機器による補助を用いずに実現できるマーカレス型 AR において，高精度なカメラポーズ推定を行うための手法が求められている．

外部機器に頼らずにカメラポーズを推定するための手法として，Absolute Pose Regression（APR）[3]を利用するものがある．APR でははじめに，RGB カメラを使って対象物体をさまざまな角度から撮影して得られた多視点画像群と，その各画像のカメラポーズのデータを用意する．それらを学習データとして深層学習を行い，完成したモデルに未知の画像を入力することによって，その画像のカメラポーズを推論によって導き出す技術である．この手法の場合，学習データの量を増やすことによってモデルの性能を上げ，カメラポーズ推定の精度を向上させることができるが，そのために手動で大量の写真を撮影することは，手間やコストの観点から現実的ではない．

以上の従来手法の内容と課題を踏まえて本研究では，手動で大量の写真を撮影する手間を増やすことなく，APR を使った AR のカメラポーズ推定の精度を従来よりも向上させることを目的とする．

2. 関連研究

2.1 Absolute Pose Regression

AR や自動運転，ロボティクスなどの分野において，カメラで撮影された画像から，撮影したカメラの 6 自由度ポーズ（位置と姿勢）を推定するタスクがあり，これを Visual Localization を呼ぶ．APR は Visual Localization に対する代表的なアプローチの一つであり，単一の RGB 画像から直接 6 自由度のカメラポーズを回帰するニューラルネット系統の総称である．

APR では，あるシーンに対して，画像とそのカメラポーズのペアを大量に与えてニューラルネットワークによる学習を行う．推論では，単一の RGB 画像 I が与えられたとき，式(1)により対応するカメラ姿勢 $\mathbf{P}$ を推定する．

$$\mathbf{P} = (\mathbf{R}, \mathbf{t}) \tag{1}$$

ここで，$\mathbf{R}$ は 3 次元回転(クォータニオンや回転行列など)，$\mathbf{t}$ は 3 次元平行移動である．

2.2 APR への AR の応用

学習させた APR モデルを，AR のリアルタイムなカメラポーズ推定へ応用した研究がある[4]．APR を採用した AR の処理の流れを図 2 に示す．この手法ではまず，対象シーンを様々な角度から撮影した多視点画像群を用意する．

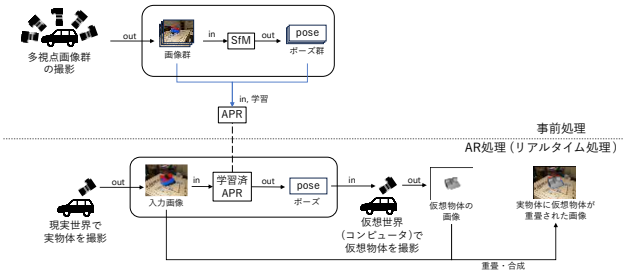

図 2　APR を用いた AR の概要

APR を AR に応用するにあたって，学習データのうち真値となるカメラポーズを画像と同時に取得することは難しい．そこで，撮影時は画像だけを取得し，その画像を元にカメラポーズを推定し，これを真値とする方法がある．ここでの推定には一般に，Structure from Motion (SfM) [5] が使われる．

3. 提案手法

3.1 提案手法の概要

APR は十分な性能を発揮するためにさまざまな視点に対応した大量の画像を入力する必要があるが，これを一般

的な RGB カメラのみで撮影して用意することは現実的ではない．そこで，任意のカメラポーズから対応する画像をレンダリングして出力することができる自由視点画像生成技術を使って， APR のデータセットを効率的に拡張する手法を提案する．なお本研究では自由視点画像生成技術として，シーンを 5 次元の放射場で表現することで高い画質を実現できる NeRF [6]を採用する．また，今後は提案手法に関連する用語を表 1 のとおりに定義して使用する．

表 1　本研究における用語とその定義

用語	定義
元画像群	学習データ用として RGB カメラで撮影された画像群
元ポーズ群	元画像群に紐づくカメラポーズ群
元データ	元画像群と元ポーズ群を合わせた学習データ
追加画像群	NeRF により生成された新規の画像群
追加ポーズ群	追加画像群に紐づくカメラポーズ群
追加データ	追加画像群と追加ポーズ群を合わせた学習データ

提案手法の処理の流れを図 3 に示す．本手法でははじめに，カメラポーズ推定を行いたいシーンの多視点画像群（元画像群）を撮影する．この時点では元画像群に対応する元ポーズ群は未知であるため，これを SfM により推定する．次に，元画像群と元ポーズ群を使って NeRF の学習を行う．学習した NeRF に，後述する手順により求めた追加ポーズ群を入力することで，対応する追加画像群が得られる．最後に，元データと追加データを合わせて使って APR の学習を行う．この APR モデルに，学習したシーンの未知の画像を入力すれば，その画像のカメラポーズを推定することができる．

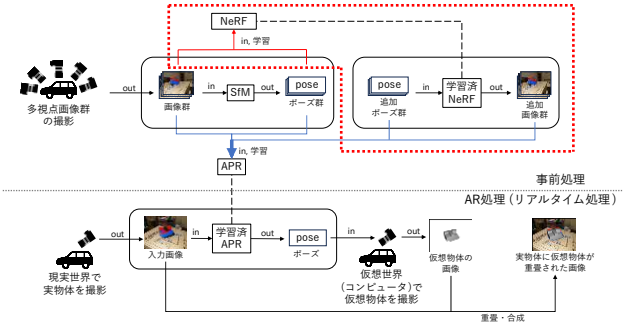

図 3　提案手法の概要

NeRF によって拡張する分のデータは，元データの 3 次元空間における各カメラ座標を元に，データの密度が低い部分を補うように決める．図 4 に，求め方の概要を示す．まず，元データの 3 次元空間を考える．図 4 (a)は，3 次元空間を原点（対象シーンの中心付近）の真上から見た図であり，便宜上 2 次元的に表現している．この 3 次元空間を，以下に示す 3 つのパラメータで分断する（図 4 (b)）．ただし図 4 において仰角方向の分断は省略している．

- 原点を基準とした仰角
- 原点を基準とした方位角
- 原点からの距離

分断して形成された各 3 次元小空間について，その中にカメラが一定数以上存在するかどうかを判別する．存在するカメラの数が既定値を下回っている小空間について，その空間内の位置に対応するカメラポーズを生成する（図 4 (c)）．このとき，生成する各カメラポーズの姿勢は，常に原点を向くようにする．以上の手順で生成されたカメラポーズを学習済みの NeRF に入力することによって，各カメラポーズでレンダリングされた新規の画像が生成され，これらの追加データを元データに加えることによって，APR の学習データがより密なものになる（図 4 (d)）．

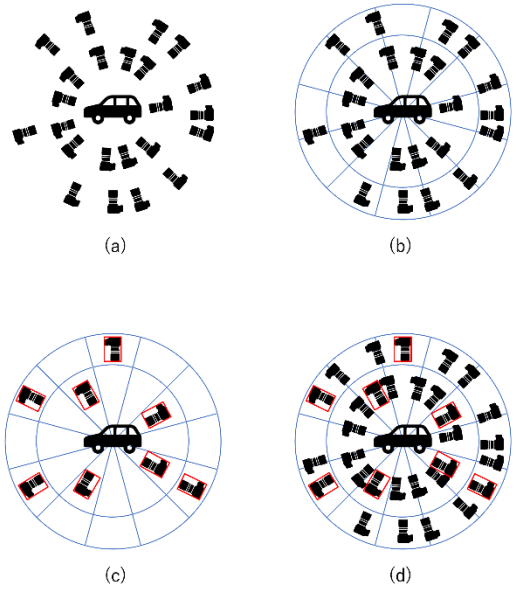

図 4　NeRF によって生成する画像のカメラポーズの決定方法

4. 実験

従来手法と提案手法の比較実験を行った．本実験では

2022 年に Chen らによって発表された高性能な APR モデルである DFNet [7]を用いた．

4.1 実験目的

本実験の目的は，本研究の提案する手法で APR のモデルを学習させることで，従来の手法で学習させるよりもカメラポーズ推定性能の高いモデルを作ることができるかどうかを確認し，提案手法の有効性を評価することである．

従来手法で APR の学習を行ったモデルと提案手法で APR の学習を行ったモデルを用意し，両モデルに対して未知である同一の画像のカメラポーズをそれぞれで推定し，2 つの推定値それぞれと真値との誤差を比較することにより，どちらのモデルの性能が優れているかを評価する．

4.2 評価指標

90 組の元データを使って DFNet で学習したモデル（本実験では従来手法モデルと呼称する）と，元データ 90 組と追加データ 92 組を合わせた合計 182 組のデータで同じく DFNet で学習したモデル（同提案手法モデルと呼称する）を用意する．未知の画像 10 枚のカメラポーズをそれぞれのモデルで推定し，その推定されたカメラポーズと，SfM であらかじめ求めていた真値との位置誤差・姿勢誤差の平均値と中央値を 2 つのモデル間で比較する．

4.3 実験手順

はじめに，NeRF および APR の学習に用いる元データを作成した．実験には，Gerrard Hall 画像データセット[8]を使用した．このデータセットはカメラポーズを含んでいないため，カメラポーズを COLMAP [9]を用いて SfM により推定した．本実験で Gerrard Hall データセットを用いたのは，このデータセットが様々なアングルと距離から撮影されたものであり，かつボケやブレ等のノイズとなる要素が少ないためである．また，COLMAP から公式に提供されているものであるため，真値となるカメラポーズを高い精度で推定しやすい．本データセットは，Gerrard Hall という屋外建造物の周りを徒歩で周回しながら撮影された，屋外シーンの写真 100 枚からなる．各画像の仕様は表 2 のとおりである．

表 2　Gerrard Hall データセットの画像の仕様

解像度	5616px×3744px
チャンネル数	3
フォーマット	JPEG

以上の手順で得られた元データを使って，NeRF の学習を行う．この処理は，nerfstudio [10]を用いて行う．

次に，従来手法と提案手法それぞれの手法で DFNet のモデルを学習させ，従来手法モデルと提案手法モデルを用意する．

続いて図 5 に示す手順によって，2 つのモデルの性能を測定する．まず，学習させたそれぞれのモデルに対して，学習に使用しなかった未知の画像を入力し，それぞれから推定ポーズを取得する．次にそれらの推定ポーズと真値との誤差を計算する．

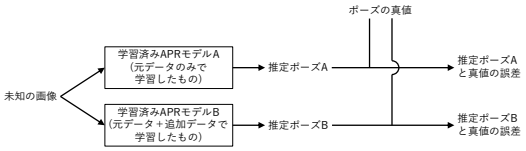

図 5　実験手順

このとき，誤差は位置誤差と姿勢誤差を個別に計算する．位置誤差は単位を[m]とし，3 次元空間上における 2 点間のユークリッド距離として求める．姿勢誤差は単位を[°]とし，次に示す手順で求める．

Step 1　真値と推定値を，それぞれクォータニオン (x, y, z, w)で$\mathbf{q}_1, \mathbf{q}_2$とする．

Step 2　$\mathbf{q}_1$と$\mathbf{q}_2$をそれぞれ単位ベクトルに正規化し，それらを$\mathbf{q}_1', \mathbf{q}_2'$とする．

Step 3　$\mathbf{q}_1'$と$\mathbf{q}_2'$の内積の絶対値をdとする．

Step 4　式(2) により，dを回転角度θに変換する．

$$\theta = 2 \times \cos^{-1} d \times \frac{180}{\pi} \tag{2}$$

以上の誤差を求める操作を 10 枚分の未知画像に対して行い，得られた 10 回分の誤差について，中央値と平均値を計算する．

4.4 実験結果と考察

はじめに，NeRF によって新たに生成された画像 92 枚のうちの 2 枚を図 6 と図 7 に示す．図 6 は，撮影対象である建物部分には大きなボケ，ブレ，オクルージョン，画としての破綻，見切れなどの学習の妨げになるようなものは見られず，学習データとして適切な画像になっていると考えられる．一方で図 7 は，画像上の建物部分の左から中心付近にかけての部分や上側の部分が，手前の草木のようなもので遮られ，オクルージョンが発生している．これは画像上の特徴点を抽出するうえで障害となる要素であり，この画像は学習データとして不適切である．

図 6　NeRF で生成された画像例 1

図 7　NeRF で生成された画像例 2

次に，従来手法モデルと提案手法モデルのカメラポーズ推定精度の比較を，中央誤差の場合と平均誤差の場合のそれぞれで表 3 および表 4 に示す．

表 3　従来手法モデルと提案手法モデルのカメラポーズ推定精度の比較（中央誤差）

	位置誤差 [m]	姿勢誤差 [°]
従来手法モデル	5.223	9.097
提案手法モデル	4.764	8.514

表 4　従来手法モデルと提案手法モデルのカメラポーズ推定精度の比較（平均誤差）

	位置誤差 [m]	姿勢誤差 [°]
従来手法モデル	5.580	9.632
提案手法モデル	5.101	9.057

表 3 に示すように，中央誤差の比較では，提案手法モデルが従来手法モデルよりも位置誤差と姿勢誤差ともに小さい値となった．また表 4 に示すように，平均誤差の比較でも同じく，提案手法モデルが従来手法モデルと比べて位置誤差と姿勢誤差の両方で従来手法モデルよりも小さい値となった．

さらに，上記のデータを元に，従来手法から提案手法へのカメラポーズ推定精度の向上率をまとめたものを表 5 に示す．

表 5　従来手法から提案手法へのカメラポーズ推定精度の向上率

	位置誤差	姿勢誤差
中央誤差	8.79%	6.41%
平均誤差	8.58%	5.97%

表 5 では，提案手法は従来手法と比較して，位置誤差は約 8.6%～8.8%，姿勢誤差は約 6.0%～6.4%優れた精度を示した．このことから，提案手法モデルは従来手法モデルと比べて，姿勢誤差よりも位置誤差でより高い改善効果が出ていることがわかる．

以上の結果から，本実験を行った条件下において，NeRF で APR 用データセットを増量することによって，APR のカメラポーズ推定精度が向上することが確認できた．しかし，提案手法モデルの結果を単体で見ると，位置誤差は約 4.8 m～5.1 m，姿勢誤差は約 8.5°～9.1°の誤差があった．これは AR で現実の物体に 3DCG モデルを重畳することを想定すると利用者がずれを感じやすいと思われ，十分な精度を達成したとは言えない．原因は複数考えられる．

1 つ目としては，元データとして使用した画像が 90 枚と少量だったことにより，学習が十分に収束しなかった可能性が考えられる．元データの量は NeRF の性能にも影響するため，これを増やすことで改善が期待できるが，手動で多視点画像を撮影できる量には現実的な限度がある．本実験では NeRF で増やすデータの数が元データの数と同程度になるように調整したが，これをさらに多くすることによって，APR 用学習データ全体の量が増え，精度が改善される可能性がある．

2 つ目は，元データのアングルが限られていたことである．Gerrard Hall データセットは建物の周りを徒歩で周回して撮影されたものであるため，カメラの高さがほぼ一定のデータであった．これにより学習データにおいて高さの変化が小さく，それが学習に影響した可能性が考えられる．ただし，実際に AR 重畳技術に APR を応用する場合，対象シーンのスケールは本実験のそれよりも小さいものになるため，高さ方向の変化は付けやすいと考えられる．

3 つ目は，NeRF で生成した画像群の中に，建物の一部を覆い隠すようにして木が映り込んでいるものがあったことである．これにより，必要な特徴点が隠れてしまったり，余計な特徴点がノイズとなってしまったりして，学習に悪影響を与えた可能性が考えられる．この問題を解決する手段としては，NeRF で新規画像をレンダリングする前に遮蔽物を除去するということが考えられる．

5. まとめ

マーカレス AR において画面内の現実の物体に対して 3DCG モデルを高精度で重ね合わせる技術を，単眼 RGB カメラのみで実現することが求められている．そのためには正確なカメラポーズをリアルタイムに推定する必要があり，これを実現する手法として APR を利用したものがあるが，その精度はまだ不十分である．

本研究では APR によるカメラポーズ推定精度を向上させるために，自由視点画像生成技術である NeRF を用いて APR の学習データを拡張する手法を提案した．

従来手法と提案手法との比較実験の結果，提案手法が位置成分で約 8.6%～8.8%，姿勢成分で約 6.0%～6.4%優れた結果を示した．また，今回の研究では，オリジナルのデータの量が少量であったことや NeRF で生成した画像の一部にオクルージョンが発生していたことなどがあり，更なる精度向上の余地があると考えられる．

今後の課題としては，NeRF で生成するデータの量をオリジナルのデータに対して多くすることや，NeRF で新規画像をレンダリングする前に遮蔽物を除去することなどが挙げられる．また，データ拡張をカメラの位置に関してではなく，カメラの姿勢に関して行うことも必要である．最終的には，完成した APR モデルを AR のシステムに組み込み，AR としての精度評価を行う予定である．

参考文献

[1] 池田聖，酒田信親，山本豪志朗：AR の教科書，マイナビ出版, 2018.
[2] Daipayan Guha, Naif M Alotaibi, Nhu Nguyen, Shaurya Gupta, Christopher McFaul, Victor X D Yang：Augmented Reality in Neurosurgery: A Review of Current Concepts and Emerging Applications, Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques, pp.235-245, 2017.
[3] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, Laura Leal-Taixe：Understanding the Limitations of CNN-based Absolute Camera Pose Regression, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:3297-3307, 2019.
[4] Changkun Liu, Yukun Zhao, Tristan Braud: KS-APR：Keyframe Selection for Robust Absolute Pose Regression, arXiv, arXiv:2308.05459, 2024
[5] Shimon Ullman：The Interpretation of Structure from Motion, Proceedings of the Royal Society of London. Series B. Biological Sciences, Vol. 203, pp.405–426 (1979).
[6] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng：NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, European Conference on Computer Vision (ECCV), 2020.
[7] Shuai Chen, Xinghui Li, Zirui Wang, Victor Adrian Prisacariu：DFNet: Enhance Absolute Pose Regression with Direct Feature Matching, European Conference on Computer Vision (ECCV), 2022.
[8] Datasets —— COLMAP 3.14.0.dev0 | 5b9a079a (2025-11-14) documentation
https://colmap.github.io/datasets.html (2026 年 1 月 29 日閲覧)
[9] COLMAP —— COLMAP 3.14.0.dev0 | 5b9a079a (2025-11-14) documentation
https://colmap.github.io/ (2026 年 1 月 29 日閲覧)
[10] GitHub - nerfstudio-project/nerfstudio: A collaboration friendly studio for NeRFs
https://github.com/nerfstudio-project/nerfstudio (2026 年 1 月 29 日閲覧)