Vol.2023-CVIM-233 No.65
2023/3/3

NeRF モデルを用いた初期値非依存なカメラポーズ推定

尾崎 優也[1,a)]　出口 大輔[1]　川西 康友[2]　村瀬 洋[1]

概要：本稿では，NeRF モデルを用いた初期値非依存なカメラポーズ推定手法を提案する．NeRF モデルを利用したカメラポーズ推定手法として inverting Neural Radiance Fields for Pose Estimation（iNeRF）がある．この iNeRF は，目標画像と推定カメラポーズからレンダリングした画像の二乗誤差の最小化に基づくカメラポーズ推定手法のため，適切な初期カメラポーズが与えられない場合は，カメラポーズを正しく推定できない．また，目標画像にオクルージョンがある場合もカメラポーズを正しく推定できない．そこでこれら 2 つの課題を解決するために，カメラポーズ回帰ネットワークを用いて iNeRF のサンプリング方法を改良した Modified iNeRF（MiNeRF）を提案する．提案手法の性能を評価するため，オクルージョンを含むさまざまなカメラポーズでレンダリングしたテスト画像を作成して評価実験を行なった．評価指標として平行移動量誤差の平均と角度誤差の平均，差分画像を用い，いずれの指標においても提案手法は誤差が減少し, その有効性を確認した．

1. はじめに

画像が撮影された位置・姿勢を求めるカメラポーズ推定の技術は，ロボットによる周囲の状況の認識や，AR における仮想情報を現実世界に合成する際などに活用される重要な技術である．このようなカメラポーズ推定手法は従来から広く研究が行なわれている．その中で，Yen-Chen ら [1] が提案した iNeRF がある．iNeRF は，NeRF の逆問題を解くことでカメラポーズ推定を行なう手法である．**図 1** に iNeRF によるカメラポーズ推定の様子を示す．学習のエポック数が進むにつれて，目標画像のカメラポーズを推定できていることが分かる．

iNeRF によるカメラポーズの推定処理は，次の（1）〜（3）の繰り返しで実現される．（1）現在の推定カメラポーズを NeRF に入力して画像をレンダリングする．（2）レンダリングした画像と目標画像の誤差を計算する．（3）誤差が小さくなるように勾配を逆伝播させてカメラポーズを更新する．ただし（2）については，画像全体で誤差を計算するには大量のメモリが必要である．そのため，一定数のピクセルをランダムに選択し，その選択したピクセルで誤差を計算することで，カメラポーズの更新を行なう．このように iNeRF は，NeRF モデルの重みの更新を行なわず，カメラポーズのみ更新を行なう手法である．

[1] 名古屋大学
Nagoya University
[2] 理化学研究所 GRP
RIKEN GRP
[a)] ozakiy@vislab.is.i.nagoya-u.ac.jp

しかし iNeRF は，（a）目標画像にオクルージョンが含まれている場合，（b）初期カメラポーズが真値のカメラポーズと大きく離れている場合，それぞれで正しくカメラポーズ推定ができない．その様子を**図 2** に示す．

実環境でのカメラポーズ推定を考えた場合，オクルージョンへの対応や，初期カメラポーズの与え方が問題となる．そこで本稿では，オクルージョンに頑健にするために，iNeRF のサンプリング方法を工夫し，カメラポーズ回帰ネットワークの学習データにデータ拡張を行なう．また，初期カメラポーズに非依存にするために，カメラポーズ回帰ネットワークによる初期値推定を行なう．これらにより，オクルージョンや適切な初期カメラポーズが与えられない場合でも動作可能なカメラポーズ推定を実現する．

2. 関連研究

2.1 NeRF

Mildenha ら [2] は，複数視点で撮影されたカメラポーズ付きの画像から，3D モデルを学習する NeRF を提案している．**図 3** に複数視点の画像から NeRF モデルを学習し，新しい視点の画像を生成する様子を示す．NeRF では，3 次元空間の位置（x,y,z）と，どの方向から見たかを表す（θ,ϕ）を入力すると，その位置の色（RGB）と密度を返す関数によって 3D モデルを記憶する．そして，ボリュームレンダリングによって画像化する．ボリュームレンダリングは微分可能レンダリングとして定式化可能なため，あるカメラポーズからレンダリングされた画像とそのカメラポーズが付与された画像の誤差を計算し，勾配を逆伝播し

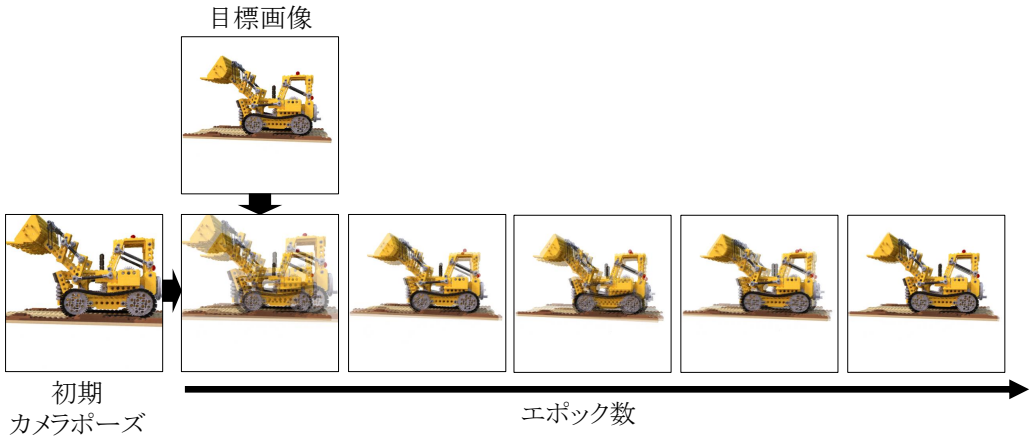

図 1 iNeRF のカメラポーズ推定例

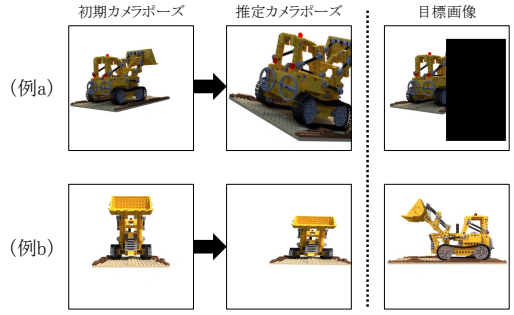

図 2 iNeRF で正しくカメラポーズ推定が不可能な例

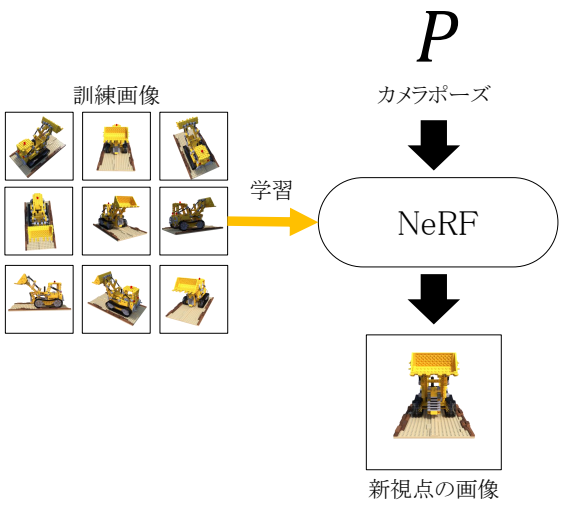

図 3 NeRF の概要

てモデルの学習が可能である．そのため，NeRF は学習やテストの際 に 3D モデルを必要とせず，カメラポーズ付きの画像のみから 3D モデルを学習することが可能である．

2.2 iNeRF の並列化による効率化

Lin ら [5] は，iNeRF のカメラポーズ推定を 1 つのカメラではなく，複数のカメラで同時に行なう方法を提案している．この研究では，NeRF モデルとして高速に学習可能な Instant NGP[3] を用いている．これにより，iNeRF が 1 つのカメラしかカメラポーズを推定できないのに対して，複数のカメラを配置して同時にカメラポーズ推定ができるようにした．

Lin らの手法 [5] は，(1) 初期カメラポーズの周囲にモンテカルロ法でカメラを複数配置し，それぞれカメラポーズを推定する．(2) その中で誤差が小さいものを一定の割合で採用する．(3) それらのカメラポーズの周囲に再び複数カメラを配置する．この (1)〜(3) を繰り返すことでカメラポーズを推定する．また，採用する割合をステップごとに少なくするといった工夫がなされている．

この手法では, 回転と平行移動を分離して推定することで効率的になること，iNeRF の損失をさまざまなものに変更して実験をし，Mean Absolute Percentage Error (MAPE) が最も良いことを示している．また，iNeRF と比較し，iNeRF でカメラポーズ推定ができない場合でもカメラポーズ推定が可能であることを報告している．しかし，初期カメラポーズやオクルージョンの影響は考慮されていない．

2.3 CNN ベースのカメラポーズ推定

Chen ら [4] は，CNN ベースのカメラポーズ回帰ネットワークを NeRF と組み合わせて学習させ，より高精度なカメラポーズ回帰ネットワークを作成する方法を提案している．この手法の概要を**図 4** に示す．この手法は，カメラポーズ回帰ネットワークが推定したカメラポーズと真値の二乗誤差 L_{gt} と，カメラポーズ回帰ネットワークが推定したカメラポーズを NeRF に入力し, レンダリングされた画像と目標画像の二乗誤差 L_{photo} の 2 つの誤差を用意し, 式 (1) のように損失を定義することで, カメラポーズがないデータでも L_{photo} を用いて学習を可能にした．

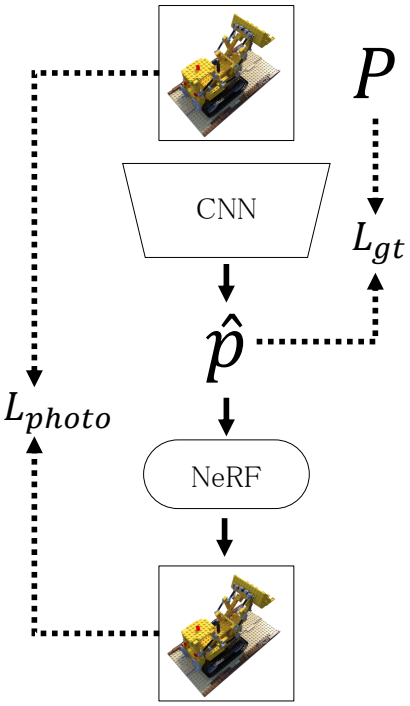

図 4 Chen らの手法の概要

$$Loss = \lambda_1 L_{photo} + \lambda_2 L_{gt} \tag{1}$$

カメラポーズ推定時には CNN ベースのポーズ回帰ネットワークのみを使用し，iNeRF のような反復的な処理を行なう必要がないため，効率的なカメラポーズ推定が可能であることを報告している．しかし，オクルージョンの影響は考慮されていない．

3. 提案手法

本稿では，カメラポーズ回帰ネットワークを用いて iNeRF によるカメラポーズ推定を行なう MiNeRF を提案する．提案手法によるカメラポーズ推定の流れを**図 5** に示す．初めに，目標画像をカメラポーズ回帰ネットワークに入力してカメラポーズを求める．そして，このカメラポーズを初期カメラポーズとして iNeRF でカメラポーズ推定を行なう．

3.1 MiNeRF

重み Θ の NeRF モデル N_Θ にカメラポーズ P を入力したときの位置 l の画素値を $N_\Theta(l|P)$，目標画像 I_t の位置 l の画素値を $I_t(l)$，エポック数を e，エポック数の最大を N，I_t の幅を W，I_t の高さを H，e エポック時のサンプリング位置の集合を $\Gamma_e \subset \{(x,y)|0 \leq x < W, 0 \leq y < H\, , x, y \in Z\}$ とすると，iNeRF は式（2）のように定式化される．

$$\widehat{P} = \underset{0 \leq e \leq N, P}{\operatorname{argmin}} \sum_{l \in \Gamma_e} |N_\Theta(l|P) - I_t(l)|^2 \tag{2}$$

iNeRF はこの Γ_e を interest region サンプリングにより

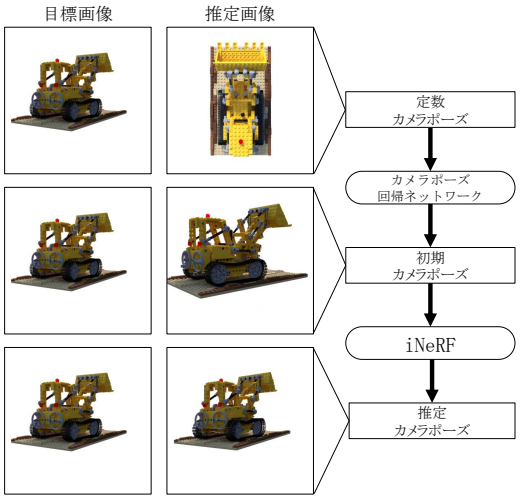

図 5 MiNeRF の概要

決定する方法を提案している．これは，画像中で重要だと判断した領域の中からランダムにサンプリングし，その範囲をエポックごとに広げ，最終的には画像全体からランダムにサンプリングを行なう方法である．しかし、オクルージョンは考慮されていない．

そのため，オクルージョンを考慮したサンプリングへと拡張する．ここでは，オクルージョンを値が 0 の矩形領域とし，ピクセルの重要度 $i_{x,y}$ $(0 \leq x < W, 0 \leq y < H)$ を式（3）のように定義する．

$$i_{x,y} = \begin{cases} 1 & \text{ピクセルの値が（0,0,0）以外} \\ 0 & \text{ピクセルの値が（0,0,0）} \end{cases} \tag{3}$$

次に，式（4）のように，$i_{x,y}$ を全体の和で除することで確率分布にする．

$$p_{x,y} = \frac{i_{x,y}}{\sum\limits_{0 \leq x < W, 0 \leq y < H} i_{x,y}} \tag{4}$$

MiNeRF では，この確率 $p_{x,y}$ に従ってサンプリングした集合を Γ_e とする．

また，iNeRF は P_0 を初期カメラポーズ，e エポック目に推定されたカメラポーズを P_e とすると，以下のように推定されたカメラポーズを反復的に利用して，次のカメラポーズを推定する．

$$\begin{aligned} P_1 &= \underset{P_0}{\operatorname{argmin}} \sum_{l \in \Gamma_1} |N_\Theta(l|P_0) - I_t(l)|^2 \\ P_2 &= \underset{P_1}{\operatorname{argmin}} \sum_{l \in \Gamma_2} |N_\Theta(l|P_1) - I_t(l)|^2 \\ &\vdots \\ P_N &= \underset{P_{N-1}}{\operatorname{argmin}} \sum_{l \in \Gamma_N} |N_\Theta(l|P_{N-1}) - I_t(l)|^2 \end{aligned} \tag{5}$$

上式の P_0 は I_t の真値のカメラポーズ P_{gt} と近いことが

前提となっている．そのため，MiNeRF では，この P_0 を 3.2 節で述べるカメラポーズ回帰ネットワーク PRN を用いて求める．

$$P_0 = PRN(I_t) \tag{6}$$

3.2 カメラポーズ回帰ネットワーク

カメラポーズ回帰ネットワーク PRN の構成を**図 6** に示す．カメラポーズ回帰ネットワークは，まず CNN ベースのネットワークで目標画像 I_t の特徴量を抽出する．その特徴量を 3 層の全結合層を通して，カメラポーズを表す 12 次元のベクトルにする．

今回は画像の特徴量を抽出するバックボーンとして ResNeSt50[6] を使用した．また，全結合層の活性化関数には ReLU を使用する．ただし，カメラポーズを表す行列は負の値を含む可能性があるため，出力層には活性化関数を使用しない．全結合層の最終出力の 12 次元のベクトルを用いて，4×4 の行列を再構成する．

このカメラポーズ回帰ネットワークの学習には，iNeRF で使用する NeRF モデルを用いて生成した画像を使用する．しかし，初期カメラポーズに非依存かつオクルージョンに頑健の両方を達成するためには，カメラポーズ回帰ネットワークもオクルージョンに頑健である必要がある．そのため，学習データに対して値が 0 の矩形領域でマスク処理をすることでデータ拡張を行なう．この矩形領域の幅，高さ，位置はいずれもランダムに決定し，さまざまなオクルージョンを学習させるためエポックごとに再生成を行なう．

最適化アルゴリズムは Adam を使用し，真値のカメラポーズ P とネットワークが推定したカメラポーズ $\widehat{P}$ の二乗誤差の和を損失関数とする．

4. 実験

カメラポーズ回帰ネットワークと iNeRF を組み合わせてカメラポーズ推定をする方法と iNeRF のサンプリングを改良する方法のそれぞれの有効性を確認する評価実験を行なった．

4.1 テストデータセットの作成

実環境に近い環境でカメラポーズ推定が可能になることを目的としていることから，データセットとして対象物体との距離が近い場合，遠い場合も含めることで，より実環境で使用される場合の条件に近づけたテストデータセットを独自に構築した．事前に構築した NeRF モデルを用い，対象物体との距離を変化させつつさまざまなカメラポーズから 300 枚の画像を生成し，基本データとする．オクルージョンの影響を評価するために，6 段階のオクルージョンレベルで，値が 0 の矩形領域をランダムな幅，高さ，位置で合成し，$300 \times 6 = 1,800$ 枚の画像を作成した．テストデータ例を図 7 に示す．

4.2 実験方法

本実験では，初期カメラポーズの影響，オクルージョンの影響の 2 つの観点で提案手法の有効性を確認する．提案手法の有効性を確認するために，以下の 3 つの手法と比較した．

比較手法 1

カメラポーズ回帰ネットワーク（PRN）のみでカメラポーズ推定を行なう．

比較手法 2

図 8 に示す 8 個のカメラポーズ P_i を用いて，$P_0 = P_i(0 \leq i \leq 7)$ として，iNeRF のみでカメラポーズ推定を行なう．

比較手法 3

カメラポーズ回帰ネットワークを利用して，$P_0 = PRN(I_t)$ とし，一様分布でサンプリングした集合を Γ_e として，iNeRF でカメラポーズ推定を行なう．

前述したテストデータセット 1,800 枚の画像に対してカメラポーズ推定を行なった結果を比較する．評価方法は，Chen ら [4] の実験で使用されている角度誤差と平行移動量誤差を用い，それぞれの誤差の平均を算出して行なった．ただし，比較手法 2 については，8 個の初期カメラポーズからカメラポーズ推定をした結果の平均を算出している．

また，目標画像と推定されたカメラポーズからレンダリングした画像の差分画像を作成し，カメラポーズ推定が正しくできているかを目視により評価した．

4.3 実験結果及び考察

平行移動量誤差の結果を**表 1**，角度誤差の結果を**表 2** に示す．各オクルージョン割合で最も誤差が小さいものを赤で示す．

実験結果から，オクルージョン割合が 0 %〜10 %から 60 %〜70 %は，角度誤差の平均，平行移動量誤差の平均ともに提案手法が最も良いことが確認できる．しかし，オクルージョン割合 0 %では比較手法 3 が最もよく，オクルージョン割合 80 %〜90 %では，角度誤差の平均が比較手法 1 が最も良いことが確認できる．

比較手法 1 と提案手法を比較すると，オクルージョン割合 80 %〜90 %を除いて，提案手法が最も良い．これは，カメラポーズ回帰ネットワークと iNeRF のカメラポーズ推定精度を比較すると，iNeRF の方が画像を比較しながらカメラポーズを推定を行えるため，カメラポーズの推定の精度が高いからだと考えられる．しかし，80 %〜90 %の角度誤差は比較手法 1 の方が良い．これは，80 %〜90 %のようにほとんどオクルージョンされているような場合は，iNeRF は比較できる部分が少ないため，カメラポーズの推定がうまくいかなくなったと考えられる．

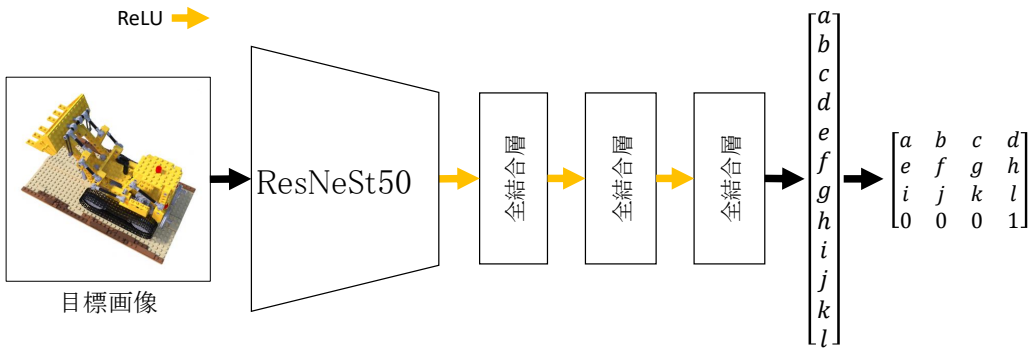

図 6 カメラポーズ回帰ネットワーク

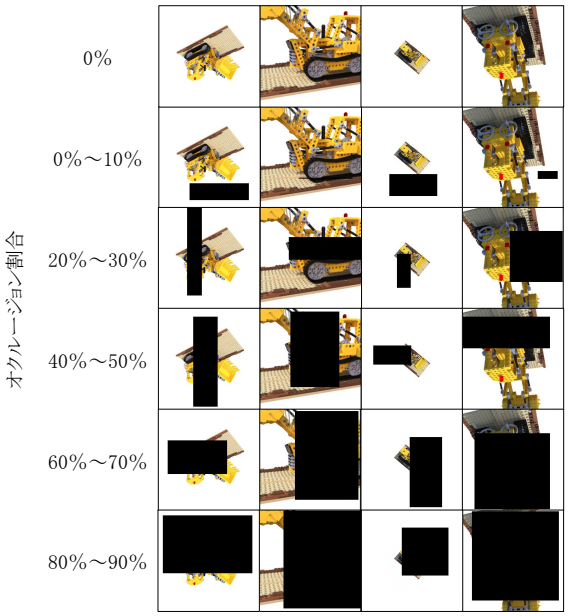

図 7 テストデータ例

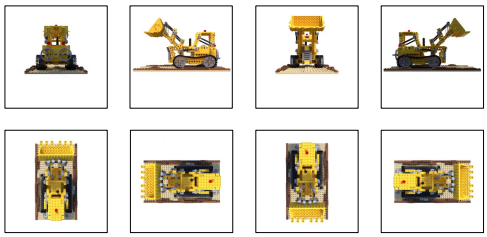

図 8 比較手法 2 の 8 個の初期カメラポーズからレンダリングした画像

比較手法 2 と提案手法を比較すると，すべてのオクルージョン割合で提案手法の方が良い．さらに，その誤差の値に注目すると比較手法 2 の平均移動誤差の平均および角度誤差の平均ともに他手法と比べて高い．これは MiNeRF が初期カメラポーズが真値のカメラポーズと離れている場合は，大きく異なる画像を比べてカメラポーズ推定を行なうため，カメラポーズ推定を正しくできないためだと考えられる．

比較手法 3 と提案手法を比較すると，オクルージョン割合 0 では比較手法 3 の方が誤差が小さい．オクルージョンがないデータでは，比較手法 3 と提案手法はサンプリングは全く同じのため，同じ結果になるはずだが誤差が生じた．この点については考慮の余地がある．また，オクルージョンがある場合は，提案手法の方が良い結果となった．MiNeRF はオクルージョンの領域以外からサンプリングを行なうため，オクルージョンがある場合でもカメラポーズ推定が行えたと考えられる．

比較手法 1 と比較手法 3 をオクルージョンがある場合で比較すると，比較手法 3 は比較手法 1 と比べてどちらの誤差も増加する傾向がある．iNeRF では，カメラポーズ推定にオクルージョンされた領域の影響を受けるため、正しくカメラポーズ推定ができない．そのため，カメラポーズ回帰ネットワークのみで求めたときのカメラポーズ推定精度よりも悪化したと考えられる．

次に，推定画像と目標画像の差分画像の結果を**図 9** に示す．この図は白いほど誤差が大きいことを表し，黒いほど誤差が小さいことを表す．この図より提案手法はすべてのオクルージョン割合で目標画像との誤差が小さくなっていることが確認できる．これはカメラポーズ回帰ネットワークのみや MiNeRF のみでカメラポーズを推定するよりも，カメラポーズ回帰ネットワークと MiNeRF を組み合わせて，カメラポーズ推定を行なった方が，より高精度なカメラポーズ推定が可能であるためだと考えられる．

最後に，推定画像の結果を**図 10** に示す．

5. まとめ

本研究では，NeRF モデルを用いた初期値非依存なカメラポーズ推定として，カメラポーズ回帰ネットワークを用いた iNeRF によるカメラポーズ推定手法を提案した．カ

表 1 平行移動量誤差の平均

オクルージョン割合	0 %	0 %〜10 %	20 %〜30 %	40 %〜50 %	60 %〜70 %	80 %〜90 %
比較手法 1	0.143	0.156	0.187	0.206	0.242	0.312
比較手法 2	6.251	6.249	6.252	6.260	6.269	6.285
比較手法 3	0.115	0.140	0.203	0.255	0.325	0.412
提案手法	0.116	0.127	0.156	0.178	0.215	0.291

表 2 角度誤差の平均

オクルージョン割合	0 5 %	0 %〜10 %	20 %〜30 %	40 %〜50 %	60 %〜70 %	80 %〜90 %
比較手法 1	1.342	1.449	1.647	1.854	2.158	3.006
比較手法 2	117.133	117.046	117.143	116.908	116.823	117.002
比較手法 3	1.044	2.648	9.872	12.891	16.621	18.953
提案手法	1.073	1.111	1.387	1.638	2.038	3.171

メラポーズ回帰ネットワークで初期カメラポーズを求めることで，初期カメラポーズに非依存なカメラポーズ推定を可能にした．また，iNeRF のサンプリング方法の改良と，カメラポーズ回帰ネットワークの学習データのデータ拡張でオクルージョンに頑健なカメラポーズ推定も可能にした．この提案手法の有効性を確認するために，さまざまなカメラポーズで，オクルージョンを含むデータセットを作成し，それを用いて評価実験を行なった．その結果，初期カメラポーズに非依存でオクルージョンに頑健な iNeRF によるカメラポーズ推定が可能であることを確認した．

参考文献

[1] Yen-Chen, L., Florence P., Barron J. T., Rodriguez A., Isola, P. and Lin T.: INeRF: Inverting Neural Radiance Fields for Pose Estimation, Proc. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1323–1330 (2021).
[2] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Proc. European Conference on Computer Vision 2020, pp.405–421 (2020).
[3] Müller, T., Evans, A., Schied, C. and Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding, ACM Transactions on Graphics, Vol.41, No.4, pp.1–15 (2022).
[4] Chen, S., Wang, Z. and Prisacariu, V.: Direct-PoseNet: Absolute pose regression with photometric consistency, Proc. 2021 International Conference on 3D Vision, pp.1175–1185 (2021).
[5] Lin, Y., Müller, T., Tremblay, J., Wen, B., Tyree, S., Evans, A., Vela, P. A. and Birchfield, S.: Parallel Inversion of Neural Radiance Fields for Robust Pose Estimation, arXiv.org, arXiv:2210.10108v1 (2022).
[6] Zhang, H., Wu, C., Zhang, Z,, Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M. and Smola, A.: ResNeSt: Split-Attention Networks, Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp.2735–2745 (2022).

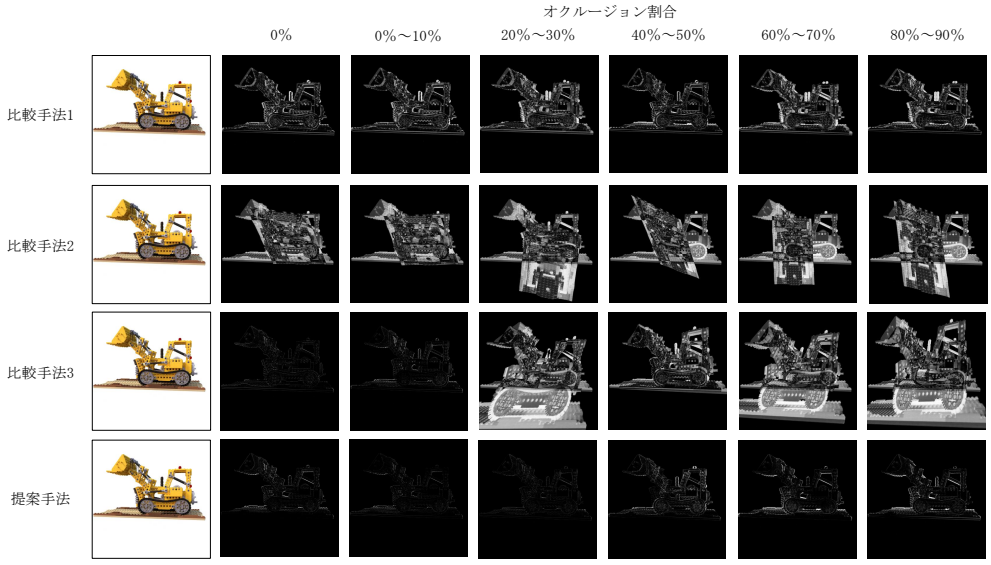

図 9 推定画像と目標画像の差分画像の結果

図 10 推定画像の結果