複数カメラを用いた３次元再構成による物体消去手法

岡田尚樹[1]　岡部誠[1]

概要：物体消去技術は、写真編集や映像制作において広く利用され、不要な物体を除去し、適切に背景を埋めることによって、あたかも最初からそこになかったかのような画像を作成することができる。近年、機械学習や深層学習の発展により、物体消去後の背景補完の精度は向上している。しかし、消去範囲が大きかったり、背景が複雑だったりすると、補完結果が不自然になるという課題がある。さらに、単一視点のみを用いる手法では、取得できる背景情報が限られるため補完の正確性に限界がある。そこで、本研究では、物体消去後の背景補完に 3 次元再構成を活用する手法を提案する。具体的には、複数のカメラを用いて、異なる視点から得られる情報を活用し、より正確な背景推定を行う。3 次元再構成には Instant-NGP を、カメラパラメータの推定には Colmap を用いることで、消去後の背景を補完する。本研究では、なるべく少ないカメラで実現することを目指し、8 台のカメラを用いた環境で実験を行う。最終的に、既存手法との比較を通じて、提案手法の有効性を検証する。

キーワード：：画像処理、深層学習、3 次元再構成

1. はじめに

物体消去技術は、写真編集や映像制作、医療画像解析など、多岐にわたる分野で活用され、その重要性を増している。写真編集の例では、不要な物体を除去し、適切に背景を補完することによって、あたかも最初からそこになかったかのような画像を作成することができる。従来の物体消去手法では、消去したい物体を手動で選択する手間や、さらに背景補完のための正解データを用意しなければならないという課題があった。

近年､機械学習や深層学習などの AI 技術の進歩により、自動化された高精度な物体認識と自然な背景補完技術が発展している[1,2,3,4]。特に敵対的生成ネットワーク（GAN）を用いた手法[5,6]や、画像生成にも使われる Stable Diffusion[7]のような拡散モデルを用いた手法[8,9]が高い性能を示している。身近な例では、スマートフォンや ChatGPT[10]などでは撮影した写真から不要な物体を自動的に消去し、背景を AI が自然に復元する技術が実装されている。しかし、消去範囲が大きかったり、背景が複雑だったりすると、補完結果が不自然になるという課題が依然として存在する。さらに、単一の視点のみでは背景情報が得られないため、背景補完の正確性には限界がある。

この問題を解決するため、山中らの折り紙の指導動画から手を消去する手法[11]がある。彼らの手法では、撮影時に片手ずつ持ち替えて撮影することで、手がない状態の被写体を学習し､背景を補完するアプローチを採用している。この手法によって、単一視点でも、被写体がどのような状態になっているかが分かるため、比較的正確な背景補完が可能となる。このように折り紙の指導動画など、物体消去後の背景が大切な場合この手法は有用である。しかし、この手法は撮影時に多くの手間と時間を要するという課題がある。そこで、我々は 3 次元再構成技術を活用し、複数視点から得られる情報を統合することで、画像中の不要な物体を消去した後の背景を高精度に推定する手法を提案する。多視点画像を用いることで、異なる視点から背景情報を取得し、従来手法のように手間をかけて繰り返し撮影を行う必要がなくなる。

本研究では、特に 3 次元再構成技術の中でも優れた性能を持つ NeRF（Neural Radiance Fields）[12]に着目し、その改良版である Instant-NGP[13]を活用する。Instant-NGP は、高精度を維持しつつ実行時間を大幅に短縮できるため、効率的な 3 次元再構成が可能である。通常、NeRF を用いた 3 次元再構成では 50～100 枚程度の画像が必要とされるが、それだけのカメラを用意するのは現実的ではない｡そこで、本研究では、なるべく少ないカメラでの実現を目指し、8 台のカメラを用いて実験を行う。8 台のカメラのみではカメラパラメータ（カメラの特性、位置、方向など）の推定が困難であるため、補助的に別の移動カメラを用いて対象物を 360° 撮影し、十分な特徴点情報を取得することで、カメラパラメータ推定を補助する。さらに、8 台のカメラを固定することで、一度パラメータを推定した後は再利用が可能となり、効率的な 3 次元再構成が実現できる。この手法により、少ないカメラ台数でも 3 次元再構成が可能となる。また、そのまま Instant-NGP を使用すると、カメラと同じ視点からレンダリングした際に手が残ってしまうという課題がある。これを解決するため、本研究では YOLOv8[14]を用いたセグメンテーションマスクを同時に Instant-NGP に入力する。これにより、手の部分を学習対象から除外し、背景のみを学習することで、レンダリング時に手が残らないようにする。

本研究は、山中らによる折り紙指導動画から手を削除する手法[11]の拡張であり、特に「手」の削除を対象とする。従来の手間を削減しつつ、より自然な背景補完を実現することを目的とし、3 次元再構成技術を活用した少ないカメラ台数での物体消去と背景補完手法を提案する。

1 静岡大学工学部数理システム工学科

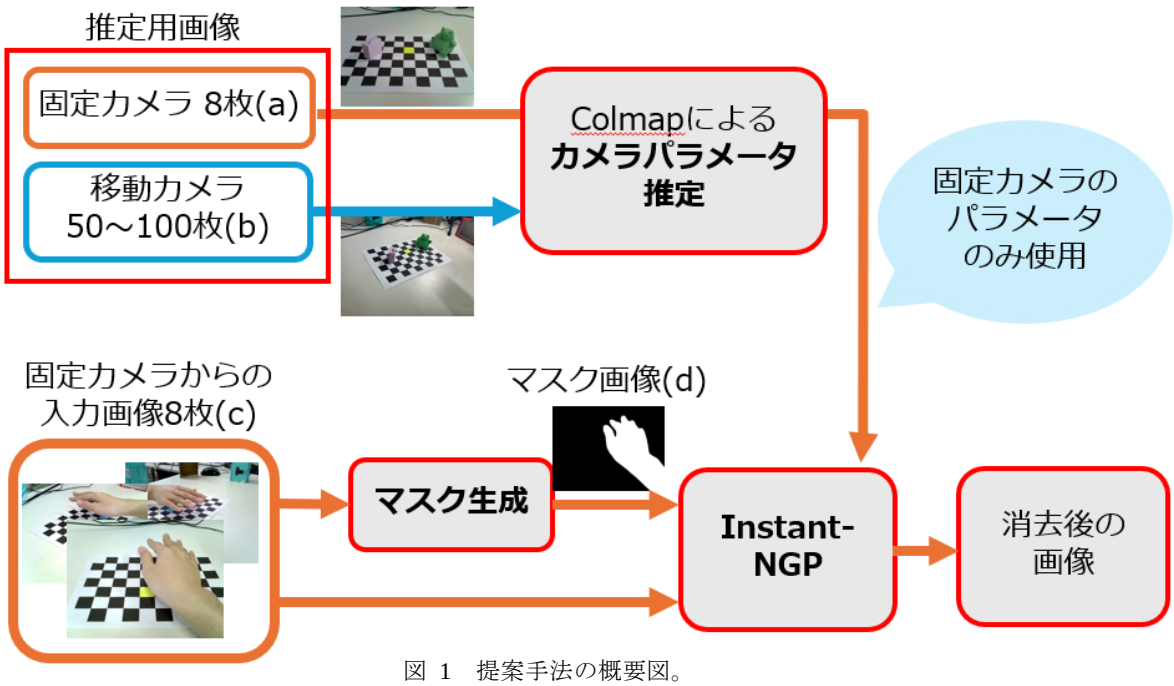

図 1　提案手法の概要図。

2. 提案手法

NeRF 技術を用いた 3 次元再構成を活用し、手の除去と背景補完を行う手法を提案する。図 1 に提案手法の概要を示す。本手法ではまず、測定用の物体を用意し、Colmap[15]を用いてカメラパラメータの推定を行う。推定したカメラパラメータのうち、移動カメラ(図 1 (b))を用いて補助的に推定したパラメータは除外し、8 台の固定カメラ(図 1(a))のパラメータのみを Instant-NGP に入力する。カメラパラメータを推定した後、固定した 8 台のカメラを移動させないようにして、対象となる画像や動画を撮影する(図 1(c))。撮影した画像と対応するセグメンテーションマスク(図 1(d))を Instant-NGP に入力し、自由視点での 3 次元再構成を行う。3 次元再構成の結果を基に、特定の視点からレンダリングを行い、最終的に手が削除された画像を生成する。

2.1 カメラパラメータ推定

Colmap は多視点画像を入力として受け取り、各入力画像の特徴点を抽出し、全ての画像ペアに対してマッチングを行うことでカメラの位置や方向を推定する。しかし、画像の数が少ないと、十分な数の特徴点とそのペアが得られず、パラメータの推定が失敗する可能性が高い。本研究では、８台のカメラのみを使用するため、このままでは推定することは困難である。そこで、この 8 台のカメラとは別のカメラを用意し、同じ状態の被写体を 360° 撮影する。50～100 枚ほど撮影し、パラメータ推定の補助として、すべての画像を Colmap に入力する。この時、測定用の物体は画像間のマッチングが行いやすいように、単調な背景や物体ではなく、色が多いものや模様のついた物体が望ましい。内部パラメータは、焦点距離や主点、レンズの歪み係数などカメラ固有のパラメータであり、同じカメラを使用する限り変化しない。カメラを固定すれば、このパラメータも変化しない。したがって、移動カメラと固定カメラの両方で撮影した画像を用いて、カメラパラメータを推定した後、固定カメラのみのパラメータを取り出す。一度推定した完了すれば、カメラを再び動かさない限りは、同じパラメータを再利用できる。

2.2 セグメンテーションマスクの生成

YOLOv8 は事前学習済みモデルが配布されており、今回のような手などの普遍的な物体であれば、特に学習データを用意せずとも推論が可能となる。撮影した 8 枚の画像を YOLOv8 に入力し、画像中の手の範囲を推定する。推定した範囲を白に、その他の背景部分を黒の二値画像に変換することで、セグメンテーションマスク(図 1(d))を生成する。

2.3 Instant-NGP による 3 次元再構成

Instant-NGP に撮影画像を入力すると、自由視点レンダリングが可能になり、視点を下に移動させると遮蔽物の下の背景が推定可能であることが確認できる。しかし、このままだでは固定カメラの視点から見たときには、遮蔽物が視界を妨げる。そのため、2.2 で生成したセグメンテーションマスクを同時に入力することで、白い部分の学習が無視され、黒い部分の学習のみが行われる。このようにすることで、手が存在しない状態の 3 次元再構成が行われて、どの視点からレンダリングを行っていても、手の下の背景が

推定された状態となる。

通常 Instant-NGP などの NeRF 技術も 50 から 100 枚程度の画像を用いて行う。そのため、8 枚のみで実行しようとすると、真ん中に物体が生成されずに、各視点の情報が分散し、不安定な結果になってしまうことも少なくない。

3. 実験結果

3.1 自由視点画像

Instant-NGP を使用して、自由視点画像を作る。自由視点画像は、ユーザーが視点を自由に変更できるため、特定の角度からだけではなく、様々な視点から対象を観察することが可能になる。図 2 は作成した自由視点画像をいくつかの視点から見たものである。 x ,y,z 軸すべての方向で範囲を狭くして、上の方に発生してしまうノイズを低減させている。入力画像 8 枚ではノイズがかなり残ってしまっているが、チェック柄があり、黄色と青色に塗られているマスがあることがわかる。このように、手の下にどのような模様があったかどうかはわかる結果となった。

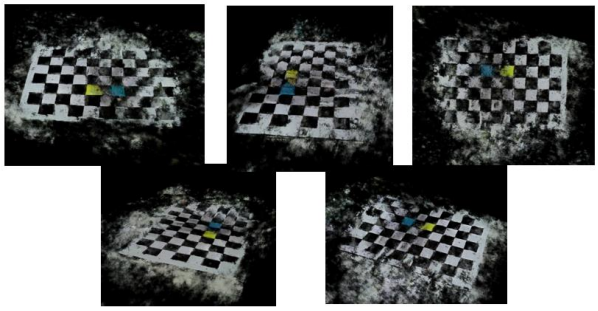

図 2　自由視点画像の結果

3.2 レンダリング結果

3.1 で生成した自由視点画像から、図 3 のように入力画像と同じ視点をレンダリングする。その後、入力画像と 2.2 で生成した入力画像のセグメンテーションマスクを用いて、入力画像の手の部分を同じ視点のレンダリング結果を補完する(図 3 左図)。図 3 を見比べると、入力画像には全く映っていなかった、手の下の青のマスが他の視点からの情報を基にしっかりと推定できていることが確認できる。チェック柄も少しずれてはいるが、手の下にチェック柄が確かにあることも確認できる。

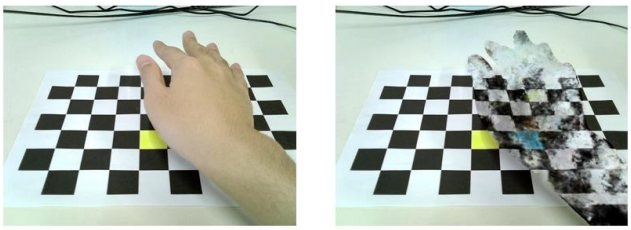

図 3　左:入力画像
右：入力画像と同じ視点のレンダリング結果

3.3 比較評価

3.2 で、物体消去の既存手法である SGEdit[9]を使用した場合を比較する(図 4)。提案手法はノイズが残ってはいるものの、手の下の模様がどの様なものであったか分かる。一方、既存手法では手があった場所の背景が推定しきれずにチェック柄が崩れてしまっている。また、3.1 の結果から確認できるように、このチェックボードには背景推定が正確に行われているか確認する為に一部のマスに色が塗ってある。既存手法は、単一視点なので白黒のチェックのみの推定しかできないが、提案手法は他の視点を用いているので、手の下にある黄色や水色が推定できていることが分かる

図 4　左：入力画像　中央：提案手法　右；既存手法

4. まとめ

本論文では、Instant-NGP を使用して、3 次元再構成を行うことで、物体消去後の背景を推定した。チェック柄の模様の推定や色の塗られているマスの復元では既存手法と比較して良い結果となった。一方で、8 台のカメラのみでは精度が十分ではなく、ノイズの発生や模様のずれ、手の形の残存といった課題が残っている。今後の展望として、現在はカメラを円形に配置してから一度も変更していないため、カメラの配置の仕方を変更したり、固定するカメラの台数を増やしたりすることで、背景補完の精度が向上することが考えられる。さらに、現状の補完結果から深層学習のノイズ除去を適用することでも精度向上が期待されるため、このような実験を行い、精度向上を目指していきたい。

参考文献

[1] Q.Dong, C.Cao, Y.Fu, “Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding”, CVPR 2022.
[2] Y.Zeng, Z.Lin, J.Yang, J.Zhang, E.Shechtman, and H.Lu, “High-Resolution Image Inpainting with Iterative Confidence Feedback and Guided Upsampling”, arXiv:2005.11742v2.
[3] Y.Zhou, C.Barnes, E.Shechtman, S.Amirghodsi, “TransFill: Reference-guided Image Inpainting by Merging Multiple Color and Spatial Transformations”, CVPR 2021.
[4] J.Yu, Z.Lin, J.Yang, X.Shen, X.Lu, T.S.Huang, “Generative Image Inpainting with Contextual Attention”, CVPR 2018.
[5] H.Liu, Z.Wan, W.Huang, Y.Song, X.Han, J.Liao, “PD-GAN:Probabilistic Diverse GAN for Image

Inpainting”, CVPR 2021.
[6] J.Yu, Z.Lin, J.Yang, X.Shen, X.Lu, T.Huang, “Free-Form Image Inpainting with Gated Convolution”, ICCV 2019.
[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022.
[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, “High-Resolution Image Synthesis With Latent Diffusion Models”, CVPR 2022.
[9] Zhiyuan Zhang, Dongdong Chen, Jing Liao, ”SGEdit: Bridging LLM with Text2Image Generative Model for Scene Graph-based Image Editing”, SIGGRAPH asia 2024.
[10] OpenAI, ChatGPT-4o, OpenAI, (オンライン), 入手先 https://openai.com/chatgpt.
[11] 山中 颯人, “手のない折り紙指導動画の作成手法”, 静岡大学大学院総合科学技術研究科修士論文, 2024.
[12] Ben Mildenhall, Pratul P.Srinivasan, Matthew Tancik, Jonathan T.Barron, Ravi Ramamoorthi, Ren Ng, “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”, ECCV 2020.
[13] Thomas Müller, Alex Evans, Christoph Schied, Alexander Keller, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding", SIGGRAPH 2022.
[14] Rejin Varghese, Sambath M, “YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness”, IEEE 2024.
[15] Johannes Lutz Schonberger, Jan-Michael Frahm, “Structure-from-motion revisited”, CVPR 2016.