令和６年度 卒業論文

複数カメラを用いた３次元再構成による物体消去手法

2025 年 2 月 13 日

静岡大学工学部 数理システム工学科

岡部研究室

50116012 岡田尚樹

目次

第1章　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。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必要があり、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。しかし、近年の 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]。例えば、スマートフォンや ChatGPT[5]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。図 1 は ChatGPT の画像処理機能を用いて、写真から緑色の物体を消去した結果を示している。確かに消去後の画像には緑色の物体は映っていないが、補完された背景にはノイズが目立ち、不自然な補完がなされていることが分かる。このように、消去する範囲が大きい場合や補完する背景が複雑な場合には、物体消去後の背景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、単一視点のみの情報では AI での復元に限界があり、物体消去後の背景情報を正確に推定できないといった問題もある。この問題を解決するために、様々な手法が提案されている。例えば、山中ら[6]は、折り紙指導動画から手を消去するために、撮影時に片手ずつ持ち替えて撮影することで、手がない状態の物体を学習し、背景を補完する手法を提案した。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。

図 1　ChatGPT による物体消去の例
右：消去前 左：消去後

そこで、本研究では 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とともに、一度の撮影のみで、手の消去を行うことを目的としている。特に、本手法は複数視点データを用い、従来の物体消去手法では困難であった複雑な背景の復元や自然な補完を実現することを目指

している。3 次元再構成とは、対象物を 360° の異なる視点から撮影し、視点間の特徴点を対応付けることで、3 次元構造を復元する技術である。本研究では、3 次元再構成の中でも特に NeRF[]を活用する。NeRF は、ニューラルネットワークを用いて 3D シーン全体を連続的な表現としてモデル化し、任意の視点からの画像を生成することが可能な技術であり、従来の 3D 再構成手法とは一線を画している。しかし、NeRF には、静的なシーンに特化していることや、学習時の計算負荷が高いといった課題がある。本研究では、8 台のカメラを固定し、カメラパラメータを一定に保つことで、フレームごとに NeRF を実行し、動的なシーンにも対応することを目指す。

また、NeRF を適用するには、カメラのレンズ特性、カメラの位置や方向などのカメラパラメータの正確な推定が不可欠である。しかし、一般的に特徴点マッチによるカメラパラメータの推定には 50~100 枚程度の画像が必要とされるため、8 台のカメラのみでは十分なデータを取得できないという問題がある。そこで、本研究では、初めに固定カメラとは別に移動可能なカメラを用いて物体の周囲を撮影し、すべてのカメラパラメータを推定した後、固定カメラのパラメータのみを抽出することで、NeRF の学習に必要なカメラ情報を確保しつつ、効率的な物体消去と背景復元を実現する。

第 2 章　関連研究

2.1　画像修復

画像修復技術とは、画像内に存在する不要な物体を削除し、削除後の背景を復元する技術である。これまで削除した後の背景補完が自然になるように様々な手法が提案されている。従来の画像修復手法には、拡散ベース手法やパッチベース手法などがある。拡散ベースの手法は削除後の欠陥領域の周囲ピクセル情報を用いて内側に拡散させることで補完を行うものである。Telea の[8]は 、この原理に基づいた手法を提案し、境界が滑らかな領域に対して一定の精度を示した。しかし、この手法では複雑なテクスチャを持つ領域では不自然な補完結果となることが多い。一方、パッチベースの手法では、画像内の類似した部分を検索し、それを欠陥領域にコピーすることで補完を行うものである。Criminisi らの[9]、特にテクスチャのある領域に対して有効であり、拡散ベースの手法よりも自然な補完が可能であるとされている。しかしながら、大きな欠損領域や背景の情報が不足している場合には、不適切なパッチが適用されることで、不自然な補間が生じるという問題がある。第 1 章で示した ChatGPT を用いた画像修復の例では、拡散ベースの手法が用いられており、物体はしっかりと削除されているものの、背景補間が不完全であることが確認された。このことからも、従来の手法では、特に広範囲の欠損領域や複雑な模様を補完する際に限界があることが分かる。

ところが、近年では、機械学習や深層学習を活用した手法が主流となっており、特に敵対的生成ネットワーク（GAN）や、画像生成にも使われる Stable Diffusion[10]のような拡散モデルを用いた手法[11]が高い性能を示している。GAN とは、生成ネットワークと識別ネットワークを競わせることで、より自然な画像を生成することが可能となる技術である。GAN を用いた手法では、Liu らの PD-GAN[12]や Yu らの[13]などがある。拡散モデルとは、画像に正規分布に従うノイズを段階的に加え、最終的に完全なノイズ状態の画像から逆拡散プロセスを通じて徐々にノイズを除去しながら画像を生成するモデルである。拡散モデルは Dhariwal[14]によって当時の最先端の生成モデルよりも優れた画像の品質を再現できることが示されており近年注目されているモデルである。この拡散モデルを物体消去に用いたものに ZHIYUAN らが提唱した SGEdit[11]がある。SGEdit は拡散モデルと大規模言語モデルを用いて、テキストで指示した画像内の物体を編集する(図 2)という手法である。図 2 からは植木鉢がきれいに消えていることが分かる。この手法は指定した物体の削除後の背景補完に拡散モデルが用いられている。一方で、このモデルに「Remove: 'hand'」と入力した結果が図 3 である。このように、SGEdit は高品質な背景補間を可能とするが、大きな欠損領域に対して正確な背景を推定するのは依然として困難であり、隠れている部分を本来のチェック柄に復元することは難しい。

本来の物体の形をどのように推定するかという問題を解決するために、今までとは異なるアプローチとして、撮影時に工夫を加えることで物体消去を可能にする手法も提案されている。例えば、山中ら[6]は、折り紙指導動画において、手を消去するために撮影時に片手ずつ持ち替えて撮影し、手がない状態の物体を学習することで、背景を補完する手法を提案した。この手法は、高精度な背景復元を可能とするが、撮影の際に手を外しながら行う必要があり、通常の撮影よりも多くの手間と時間を要するという課題がある。

このように、従来の拡散ベースやパッチベースの手法では、大きな欠損領域の補完が困難であり、GAN や拡散モデルを活用した手法でも、隠れている領域の背景を正確に推定することが難しいという課題がある。また、撮影時に工夫を行う手法も提案されているが、撮影の手間が大きく増えすぎるという問題がある。そこで、本研究では、3 次元再構成技術（2.3 節で述べる）を活用し、複数視点から得られる情報を統合することで、より高精度な背景推定を行うことを目指す。

図 2　SGEdit に物体消去の成功例
右：変更前 左：変更後

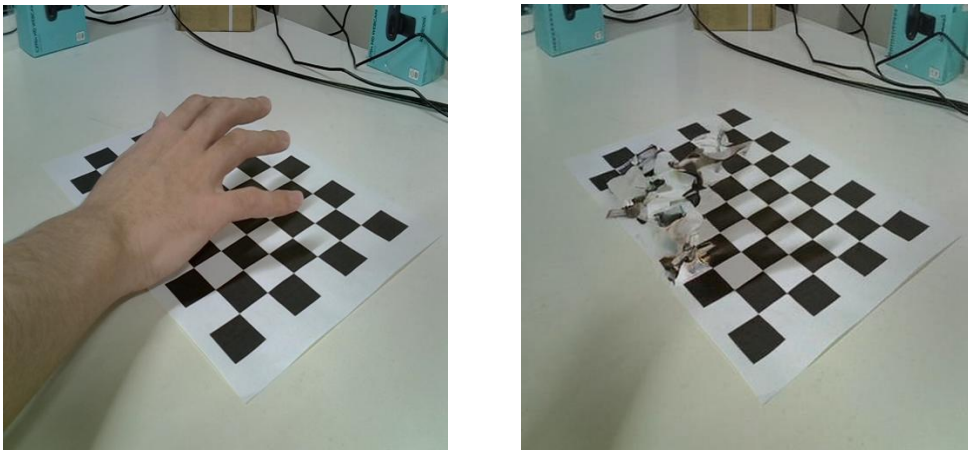

図 3　SGEdit による物体削除の限界例
右：消去前 左：消去後

2.2　セグメンテーション技術

セグメンテーションとは、画像を意味のある領域に分割することであり、物体検出や画像解析において広く活用されている。特に、物体消去においては、削除対象となる物体を正確に認識し、その領域をマスクとして取得することが重要である。適切なセグメンテーションが行われなければ、物体消去後の背景推定の精度が低下し、不自然な結果を生じる可能性がある。セグメンテーション技術には深層学習を活用した多くの手法が提案されている。代表的な手法として、U-Net[15]や Mask R-CNN[16]などの畳み込みニューラルネットワーク(CNN)ベースの手法、物体検出に多く使用される Yolov8[17]などが挙げられる。U-Net や Mask R-CNN などは医療画像解析や精細な物体領域の分割が求められる応用で高い性能を示しており、ピクセル単位のセグメンテーションが可能である。しかし、これらのモデルは学習に大量のデータセットを必要とするため、特定の物体や状況に特化したモデルを構築する際には、データ収集とアノテーション（ラベル付け）の負担が大きいという課題がある。また、モデルの構造が複雑であるため、リアルタイム処理には不向きである点も課題の一つである。これに対し、YOLOv8（You Only Look Once version 8）は事前学習済みのモデルが提供されており、人物、車、動物などの普遍的な物体であれば、学習データを用意することなく、高速で高精度に検出できる。また、必要に応じて事前学習済みモデルにファインチューニングを行うことで、少ないデータセットでも特定の物体に特化した検出が可能となる。これにより、データ収集やアノテーション作業の負担を軽減しつつ、高い精度のセグメンテーションを実現できる点が大きな利点である。以上のことから本研究では、削除対象を手に絞ることで、YOLOv8 の事前学習済みモデルだけで削除対象物のセグメンテーションマスクの推定を行う。

2.3　3 次元再構成

本研究では、NeRF を用いて画像内の物体を消去した後の背景を推定する。NeRF（Neural Radiance Fields）は Mildenhall ら[7]によって提案された手法であり、ニューラルネットワークを用いて 3 次元空間上の各ピクセルの RGB 値や物体の透明度を学習する。そして、ボリュームレンダリングによって、複数視点から撮影された画像を基に、新しい視点からの画像を生成することができる。ボリュームレンダリングとは、3 次元空間を通る光の線をシミュレーションし、ニューラルネットワークの出力からその光が通過する物体の密度や色を計算し、最終的な画像を合成する技術である。このようにして、360°から撮影された画像を基に、物体を削除した後の背景を他のカメラ視点から推定することが可能となり、これにより正確かつ自然な背景補間が実現されている。NeRF は提案以降、いくつかの改良が加えられ、その中でも代表的なものとして Instant-NGP が挙げられる。Instant-NGP は、Müller ら[19]によって提案された手法で、NeRF の計算時間を大幅に削減し、数十秒で 3 次元再構成を行うことを可能にしている。従来の NeRF では、計算負荷が高いために数時間を要する場合が多かったが、Instant-NGP では効率的なデータ構造と計算アルゴリズムを採用することで、実用的なスピードを実現している。本研究では Instant-NGP を用いて 3 次元再構成を高速に行う。

2.4　カメラパラメータ推定

NeRF の実行には、カメラのレンズ特性、カメラの位置や方向などのカメラパラメータの正確な推定が必要である。カメラパラメータには、焦点距離や主点、レンズの歪み係数といった内部パラメータと、カメラの位置や向きを示す外部パラメータがある。これらのパラメータが正確に求められない場合、NeRF による 3 次元再構成の精度が大幅に低下し、不自然なレンダリング結果を生じる可能性がある。

本研究では、多くの研究でカメラパラメータの推定に使用される colmap を用いた。Colmap は、Structure from Motion（SfM）[19]と Multi-View Stereo（MVS）[20]を統合したオープンソースの 3D 再構成ツールであり、複数視点の画像からカメラパラメータを高精度に推定することができる。まず、画像内の特徴点を検出し、異なる視点の画像間で対応する特徴点をマッチングする。次に、特徴点の対応情報をもとに、カメラの初期位置を推定する。Colmap では、Structure from Motion の手法を用いて、初期のカメラ位置と 3D 点を計算し、その後、順次新しい画像を追加しながらカメラの位置と向きを最適化していく。

本研究では、同じ種類のカメラを使用し、カメラの位置を固定することで、一度 Colmap を用いてカメラパラメータを推定した後は、そのパラメータを使い回すことができる。

第 3 章　提案手法

本研究は、山中らの折り紙のお手本動画から手を削除する研究[6]を拡張した研究である。そのため、本研究では特に「手」の削除を対象とする。NeRF 技術を活用して手の除去を行い、背景情報を補完する手法を提案する(図 4)。

まず、カメラパラメータの推定を行うために、測定用の物体を用意し、固定カメラで物体を撮影する(図 4(a))。その後、同じ物体を移動させずに異なるカメラで 360° 囲むようにして、撮影を行う(図 4(b))。得られた多視点画像から Colmap を用いてカメラパラメータを推定し、推定の補助に使用したカメラの情報固定カメラのパラメータのみを使用する。次に、固定した 8 台のカメラで目的の写真や動画を撮影する(図 4(c))。撮影した画像を YOLOv8 に入力し、手と背景をセグメンテーションする。その後、セグメンテーション結果を基に白黒のマスク画像を作成する(図 4(d))。最後に、撮影画像と対応するマスク画像を Instant-NGP に入力し、自由視点での 3 次元再構成を行う。最終的に、指定した視点からのレンダリングを行い、手が除去された画像を得る。

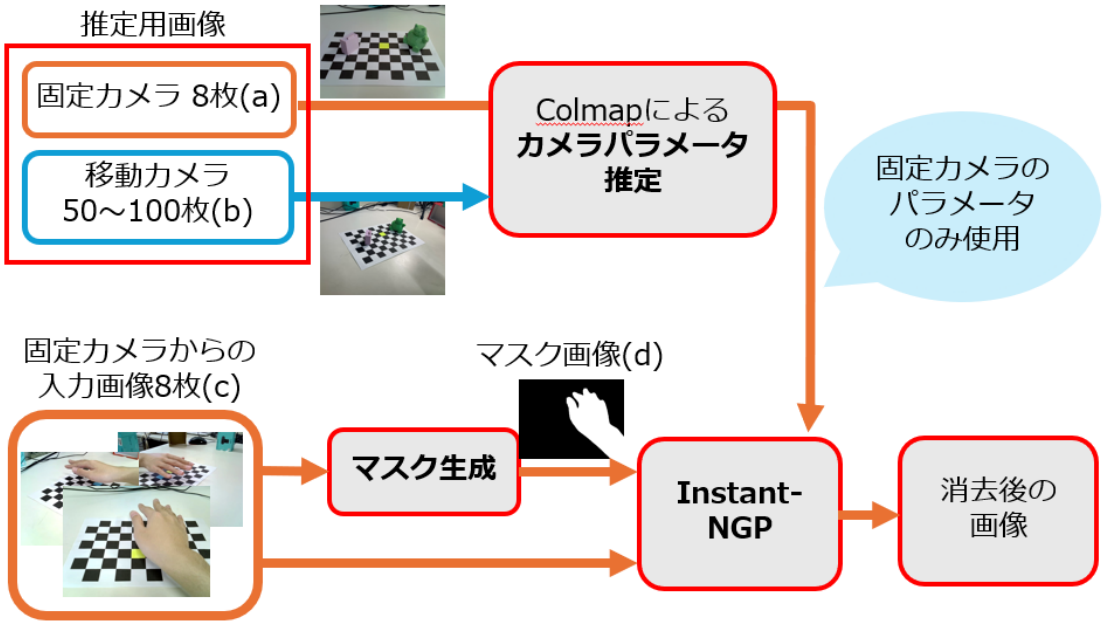

図 4　提案手法の概要図

3.1　データ収集

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。カメラは物体を囲むようにして円形

に 8 方向に配置した。その 8 個のカメラで同時に撮影し、8 つの視点画像から NeRF を実行し、背景補完を行う。しかし、8 枚だけでは、特徴点を十分に得ることができず 3.2 に述べるカメラパラメータの推定ができない。そのため、固定カメラとは別に図 5 のように、物体の周り 360° 囲むようにして撮影する。図 5 は周りから撮影している様子を可視化したもので、図 5 の赤四角は画像がどのように見えているかを示している。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(図 6)。

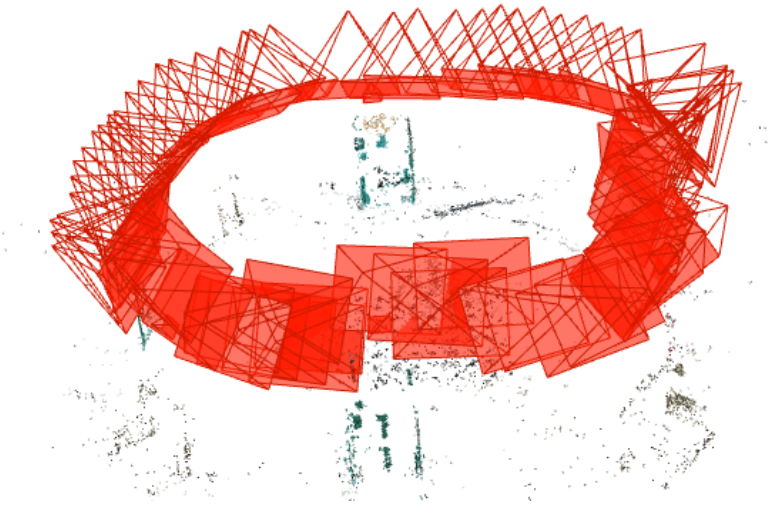

図 5　撮影位置図例

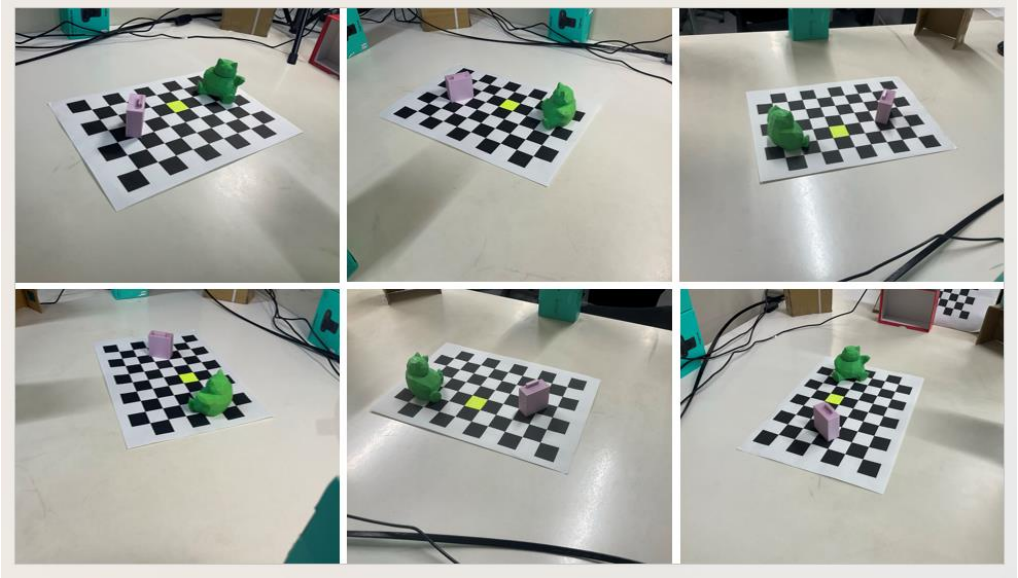

図 6　撮影例

3.2　カメラパラメータ推定

3.1 で集めた多視点画像から、Colmap を用いて特徴点抽出を行う。その結果を基に特徴点マッチングを行う。これらの操作は、Colmap のデフォルトの値を用いた。前章でも述べたように、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カメラパラメータ推定を補助する。

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でカメラ固有のパラメータであり、同じカメラを用いれば、変わることはない。外部パラメータはカメラの位置や向きを示し、カメラが動かないように固定してあれば、このパラメータも変わることはない。したがって、移動カメラと固定カメラの両方で撮った多視点画像から、前述のようにし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ことで、一度推定した後は、カメラを再び動かさない限りは、カメラパラメータを再利用できる。

3.3　セグメンテーションマスクの生成

カメラパラメータを推定した後、目的の画像や動画を撮る。

YOLOv8 は事前学習モデルが配布されており、高速かつ高性能にセグメンテーションが行うことができる。事前学習済みモデルによって、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。図 7 右図にその結果を示す。この結果を用いて図 7 左図のように、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。

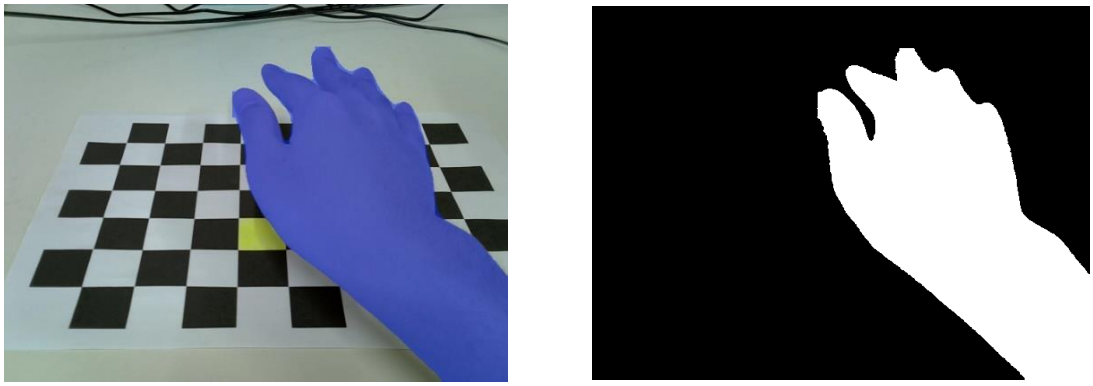

図 7　右：YOLOv8 によるセグメンテーション結果
左：セグメンテーションマスク

3.4 Instant-NGP による 3 次元再構成

3.4.1 自由視点の生成

3.2 で推定した 8 台のカメラパラメータ、入力画像 8 枚を Instant-NGP に入力する。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手などの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、3.3 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。

3.4.2 レンダリング

自由視点を作成した後、取り出したい視点のカメラパラメータを入力しレンダリングを行うことで、任意視点から見た結果が出力できる。今回は入力画像と同じ視点を用いれば良いため、入力した 8 台のカメラのパラメータのうち 1 つを選んで、レンダリングを行う。

第 4 章　結果

4.1　Instant-NGP による 3 次元再構成の結果

4.1.1　自由視点

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 8 は Instant-NGP に入力した画像である。固定した 8 台の固定カメラで、様々な方向から対象物を撮影する。この 8 枚の入力画像に加え、手の部分にセグメンテーションマスクを取った計 16 枚で 3 次元再構成を行う。図 9 がその結果をいくつかの視点からみたものをである。x ,y,z 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る。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分かる結果となった。

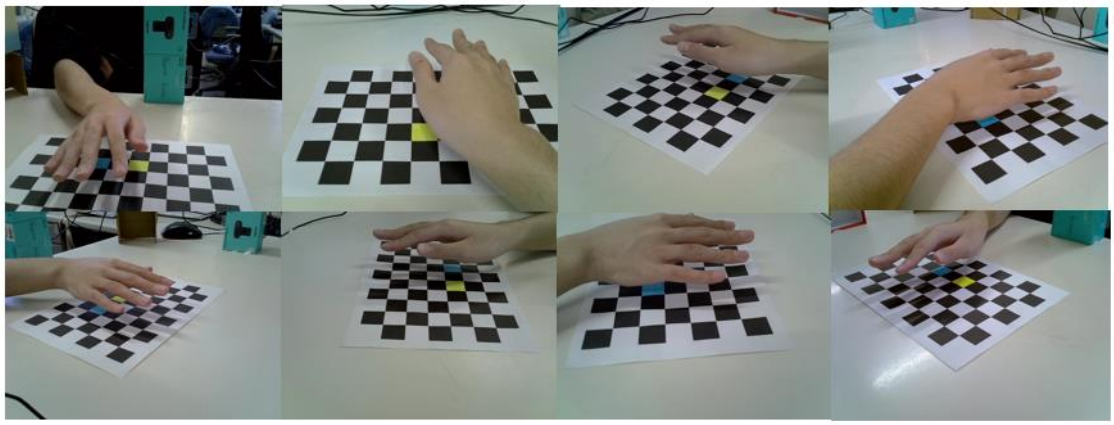

図 8　8 台の固定カメラによる入力画像

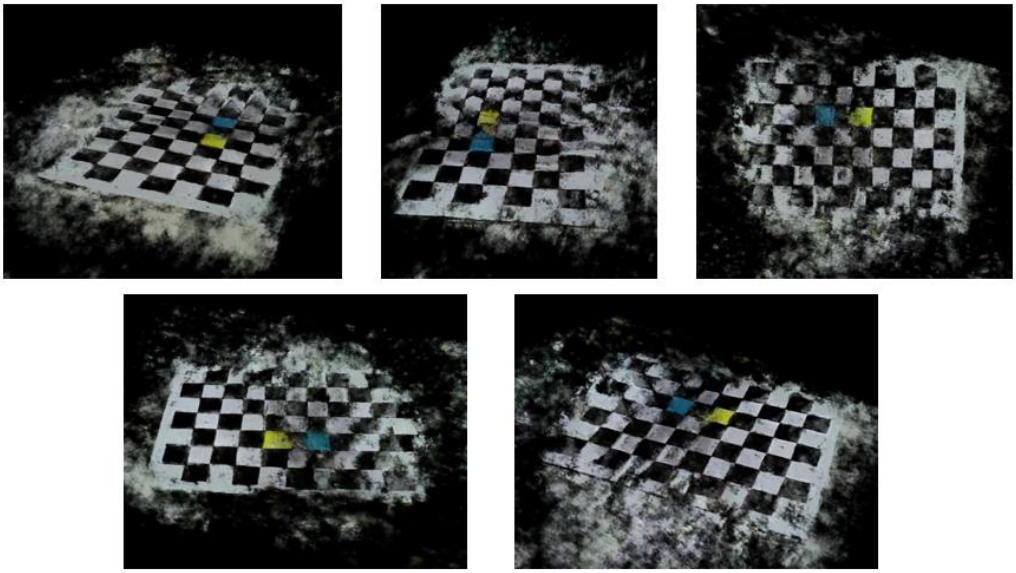

図 9　Instant-NGP の自由視点結果

4.1.2　レンダリング結果

4.1.1 で生成した自由視点画像から図 10 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 3.3 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 10 左図)。図 10 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。チェック柄において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。

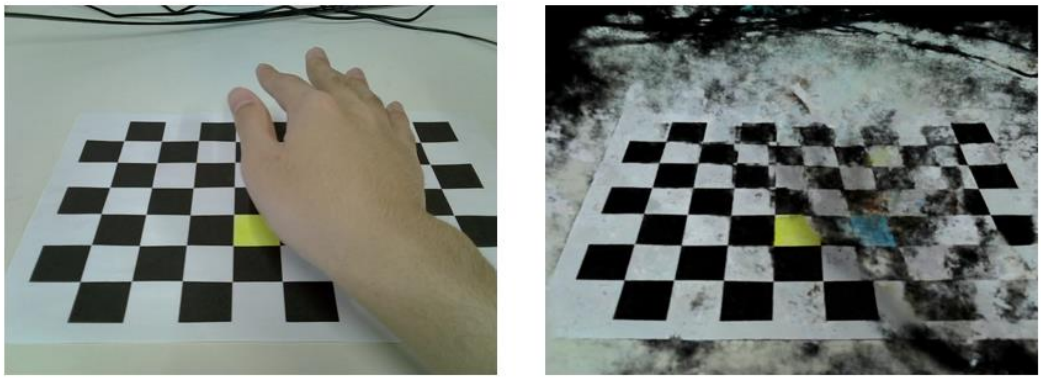

図 10 左:入力画像 右：入力画像と同じ視点のレンダリング結果

4.2　比較評価

4.1.1　視覚評価

4.1.2 で生成したレンダリング結果をもとに入力画像の手を置き換えて、物体消去の既存手法である SGEdit[11]を使用した場合を比較する(図 10)。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、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、手の下にある水色が推定できていることが分かる。

図 10 既存手法と提案手法の結果比較
上：入力画像 左：既存手法　右：提案手法

4.2.2　実行時間の比較

提案手法と SGEdit による手の削除を行った場合の提案手法 では、まずカメラパラメータ推定のためのデータ収集（撮影）に約 5 分 を要し、その後、Colmap による特徴点の抽出および特徴点マッチングが 約 1 分、特徴点を用いたカメラパラメータ推定に 約 2 分 かかる。これらを合わせると、3 次元再構成の前処理に約 10 分 を要する。さらに、YOLOv8 を用いたセグメンテーションマスクの生成は 数十秒 で完了し、Instant-NGP を用いた 3 次元再構成の処理は 最長でも 1 分程度 で収束する。そのため、全体の処理時間は約 15 分で完了 する。

一方で、既存手法（SGEdit） は、拡散モデルを用いて手の削除を行うため、処理に 約 2 時間 30 分 を要する。拡散モデルでは、画像生成時にノイズを徐々に除去しながら多数のステップを経る ため、各ステップでの計算コストが高く、全体の処理時間が長くなる傾向がある。

この結果から、提案手法は既存手法と比較して大幅に処理時間を短縮できることが確認された。

	実行時間
SGEdit	約 2 時間 30 分
提案手法	約 15 分

表 1　撮影してから、削除結果が出るまでの時間の比較

第 5 章　まとめと今後の展望

本研究では，NeRF 技術を用いて、多視点画像から手の削除を行った。NeRF は本来数十枚から数百枚を用いて行うものなので、今回の 8 台のみの結果ではノイズがかなり残ってしまった。しかし、既存手法では、チェック柄ができていなかった部分や色が塗られているマスの推定は難しかったのに対して、提案手法では、ノイズやズレは少なからず折るものの、手の下にチェック柄があることや、どのマスに色が塗られているかが分かる結果となった。今後の展望として、カメラの配置の仕方やカメラの台数を増やすことで精度が向上することが考えられる。今回、カメラを手に対して少し高い位置にして、8 台とも同じ高さで円形に配置してから、一度も変更していないので、様々な配置を試して見ることで、結果が変わるかもしれない。配置する円形の半径を狭くして、ある視点から見える物体が、別の視点でも観測できる領域を多くすることで、結果が少しよくなるのではないかと期待される。さらに、8 台という少ないカメラで行っているので、精度が芳しくなくノイズが多くなっているので、手軽さの範囲を超えない範囲で、もう少しカメラを増やすことで、結果の安定性を高められるかもしれない。

また、結果の安定性を高めるのは別に、今の図 10 のような結果からノイズや不整合性を減らしていく方法も考えられる。例えば、深層学習のノイズ除去を行う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。

謝辞

本研究及び論文の作成にあたり，研究の着想や論文執筆等，多くのご指導，ご助言を頂きました静岡大学工学部の岡部誠准教授に心から感謝申し上げます．また，ご助力頂いた修士課程学生及び学部生の皆様に深く感謝致します．

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 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] OpenAI, ChatGPT-4o, OpenAI, 入手先 https://openai.com/chatgpt.

[6] 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.

[7] 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.

[8] Alexandru Telea, “An Image Inpainting Technique Based on the Fast Marching Method”, Journal of Graphics Tools 2004.

[9] A. Criminisi, P. Perez, K. Toyama, “Region filling and object removal by exemplar-based image inpainting”, IEEE 2004.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022.

[11] Zhiyuan Zhang, Dongdong Chen, Jing Liao, ”SGEdit: Bridging LLM with Text2Image Generative Model for Scene Graph-based Image Editing”, SIGGRAPH asia 2024.

[12] H.Liu, Z.Wan, W.Huang, Y.Song, X.Han, J.Liao, “PD-GAN:Probabilistic Diverse GAN for Image Inpainting”, CVPR 2021.

[13] J .Yu, Z.Lin, J.Yang, X.Shen, X.Lu, T.Huang, “Free-Form Image Inpainting Convolution”, ICCV 2019.

[14] Prafulla Dhariwal, “Diffusion Models Beat GANs on Image Synthesis”, arXiv:2105.05233v4.

[15] Olaf Ronneberger, Philipp Fischer, Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, MICCAI 2015.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, "Mask R-CNN”, ICCV 2017.

[17] Rejin Varghese, Sambath M, “YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness”, IEEE 2024.

[18] Thomas Müller, Alex Evans, Christoph Schied, Alexander Keller, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding", SIGGRAPH 2022.

[19] Johannes Lutz Schonberger, Jan-Michael Frahm, “Structure-from-motion revisited”, CVPR 2016.

[20] Johannes L. Schönberger, Enliang Zheng, Marc Pollefeys, Jan-Michael Frahm, “Pixelwise View Selection for Unstructured Multi-View Stereo”, ECCV 2016.