複数視点による三次元再構成を用いた物体消去

岡部研究室　　525C5008　　岡田尚樹

1．　はじめに

近年，画像中から不要な物体を消去する技術は，写真編集やSNS，映像制作などの身近な場面で広く利用されている。例えば，観光地で撮影した写真に写り込んだ通行人を消去したり，電線や看板などの不要物を取り除いたりすることで，見栄えの良い画像を簡単に得ることが可能となっている。このような物体消去技術は，近年の深層学習をはじめとする AI 技術の発展により大きく進歩した。特に画像生成分野では，周囲の風景や文脈を考慮し，一見すると初めからその物体が存在しなかったかのような，自然な補完結果を生成できる手法が多数提案されている。一方で，既存の AI ベースの物体消去手法にはいくつかの制限が存在する。まず，消去範囲が大きい場合には，周囲の情報だけでは十分な補完が困難となり，不自然な結果が生じやすい。また，周囲に利用可能な情報が少ない状況や，複雑な模様・構造を持つ物体を再現する場合には，正確な復元が難しいという問題がある。さらに，本来存在しない関係のない物体が生成されてしまうなど，意図しない生成結果が現れることも少なくない。これらの問題の根本的な要因として，AI による生成結果は学習データに基づく推定であり，あくまで「想像」によって補完が行われている点が挙げられる。そのため，視覚的には自然に見える場合であっても，実際のシーンと整合しない構造や模様が生成されることがあり，正確性という観点では限界がある。つまり，既存手法の多くは，存在しない情報を生成によって補完する点に限界を抱えている。

そこで本研究では，生成に依存するのではなく，消去領域に対応する実在の情報を別視点から取得するという発想に基づき，消去対象を含む画像とは別に，同一シーンを異なる視点（例えば横方向からの視点）で撮影した画像を利用する手法を提案する。既存の画像生成モデルに同一シーンの視点を入力とし，物体消去を行わせる場合（4.2 比較結果参照），視点間での一貫性を保つことは依然として困難であり，現状のモデルでは同一シーンとして整合した結果を得ることが難しい。一方で，本研究で提案する手法では，同一シーンの複数視点画像を三次元再構成することで，シーン全体の幾何構造を明示的に扱うことが可能となる。これにより，視点間での不整合や模様の破綻を抑えた，一貫性のある物体消去を実現する。さらに，提案手法は既存の生成モデルに対する再学習や追加学習を一切必要としないため，低コストかつ導入の手間が少ない形で高精度な物体消去を実現できる点に特徴がある。

2．　関連研究

2.1. 三次元再構成（MASt3R）

複数の画像から三次元構造を復元する手法として， 従来の複雑な処理手順を省略し， 画像から直接三次元情報を推定する MAST3R[1]が提案されている。 これらの手法は，従来用いられてきた特徴点検出や幾何計算を段階的に行う処理に代わり，大量のデータから学習したモデルを用いて，画像から三次元構造を直接推定する点に特徴がある。

2 枚の入力画像を用いて，各画素が空間中のどの位置を見ているかを学習に基づいて推定し，シーンの三次元構造を点群として出力する。画像間の関係を同時に考慮しながら三次元位置を予測することで，事前にカメラの情報が与えられていない場合であっても，画像のみから三次元構造を復元できる。

また，前述の三次元構造の推定と同時に，画像間で同じ場所を写している画素同士を対応付ける情報を学習することで，三次元的な位置関係と画素間の対応関係の両方を考慮した推定が可能となっている。これにより，見た目が似ているだけの異なる場所に誤って対応付けられることを抑え，大きな視点変化が存在する場合であっても，安定した対応点推定を実現している。さらに，多数の画素を扱う場合においても，学習結果を活かした効率的な対応点探索が導入されている。

2.2. セグメンテーションマスク

画像中の特定の物体領域を抽出する手法として，物体検知およびセグメンテーションに基づくアプローチが広く用いられている。これらの手法は，画像中に存在する物体の位置や範囲を自動的に推定することを目的としており，前処理や対象領域の特定において重要な役割を果たす。YOLO は 1 回の推論で物体の位置を推定できる高速な物体検出手法であり，追加学習や実装の容易さから，多様な応用において物体領域抽出のための手法として広く利用されている。

2.3. 画像生成

画像中の欠損領域を自然に補完する手法として，拡散モデルに基づくインペイント技術が広く研究されている。ノイズを段階的に除去する過程を通じて画像を生成する枠組みを用いることで，周囲のピクセルや文章を考慮した高品質な補完結果を得られる点に特徴がある。視覚的に自然な画像を得ることが可能である。そのため，画像修復や画像編集において，高い表現力を示す手法として利用されている。

近年では，Gemini に代表される大規模な画像生成モデルが提案されており，テキスト入力や画像入力を組み合わせて用いることで，より柔軟な画像生成や編集が可能となっている。これらのモデルは，テキストによる指示に基づく画像生成に加え，既存画像を入力として与えた上で，画像の編集を行う用途にも利用されている。

3．　提案手法

本研究では，多視点画像を用いた三次元再構成に基づき，三次元空間上で物体消去を行った後，画像生成モデルを用いた二次元画像処理によって最終的な画像品質を向上させる手法を提案する。

3.1. 三次元処理

まず，消去対象となる物体領域を指定するため，入力画像に対して YOLO を用いたセグメンテーションを行い，消去対象のマスクを生成する。次に，MASt3R を用いて複数視点画像間の対応付けおよび三次元再構成を行うが，この段階ではマスクを適用せず，消去対象を含めた状態で画像間の対応付けを完了させる。対応付けの段階であらかじめマスクを適用すると，利用可能な画素数が減少し，画素のペアが十分でなくなるため，入力画像に写っているにもかかわらず，再構成できない領域が生じる可能性がある。（図 1）

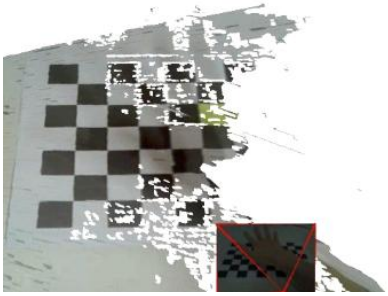

図 1 ：再構成に失敗した例(赤枠が入力画像)

そのため，本研究では，まず十分な対応付けを行った後，三次元点群の生成段階において，マスクで指定された領域に由来する点群のみを除外する。これにより，三次元空間上での物体消去を実現する。物体消去後の背景については，別視点から観測された情報が存在するため，それらを用いて自然に補完される。このため，あたかも初めから物体が存在しなかったかのような結果を得ることができる。

また，視点数が少ない場合には，再構成された三次元点群において，図 2 に示すような線状のノイズが発生することがある。これに対し，本研究では，三次元点群が最も密に分布する方向を平面として推定し，その平面から一定距離以上離れた点を除去することで，不要なノイズを排除している。

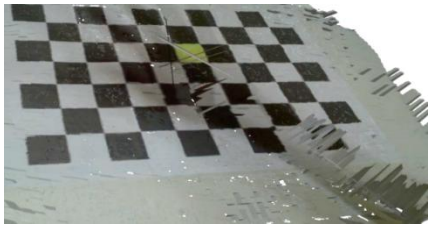

図 2：線状のノイズ(中央から右下)

得られた三次元再構成結果を基に，入力画像と同一の視点位置からシーンを再レンダリングする。次に，消去対象マスクを再び利用し，入力画像のうちマスクで指定された領域に対してのみ，再レンダリング画像を挿入する。(図 3 左)これにより，消去対象領域以外の画素は入力画像の情報を保持したまま，三次元再構成に基づく背景補完を行うことができる。しかし，三次元再構成の精度には限界があるため，この時点では入力画像と完全に一致する高品質な画像を得ることは困難である。そこで，最終的な画像品質を向上させるため，入力画像と再レンダリング画像の最適化処理を行う。

3.2. 最適化処理

再レンダリング画像は入力画像と完全には一致せず，欠損部分や境界付近に違和感が生じる場合がある。そこで二次元画像処理および画像生成モデルを用いて，最終的な画像品質の向上を図る。まず，ポアソンブレンディングを適用し，再レンダリング画像と入力画像の境界を滑らかに接続する。(図 3 右)これにより，後段の画像修復処理において，消去領域の境界が強調されることを抑制する。最後に，Gemini の画像生成機能を用いて，周囲の画素と消去後の領域が調和するようにノイズ除去および微小な欠損領域の補完を行い，視覚的に自然な最終結果を得る。

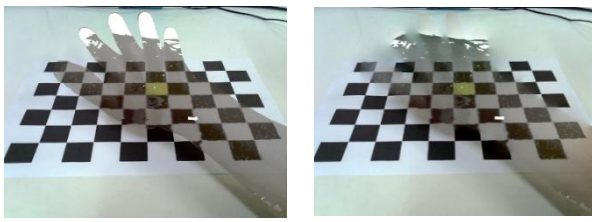

図 3：入力画像に挿入(左)　　ポアソンブレンディング後(右)

4．　実験結果

本研究では、実験に消去対象として手を，背景としてチェックボード模様を用いた。手は日常環境における普遍的な物体であり，YOLO による検出およびセグメンテーションが比較的安定して行える対象である。また，背景には規則的かつ複雑な模様を持つチェックボードを用いた。チェックボードは，連続性や規則性が明確であるため，物体消去後における模様の再現性や構造の破綻を視覚的に確認しやすいという利点がある。さらに，本実験で用いたチェックボードには，白黒の格子に加えて黄色に塗られた領域が含まれている。これは，学習データ中に存在しない可能性のある色や模様に対しても，背景を適切に復元できるかを検証するためである。

4.1. MAST3R 上で物体消去を行った結果

まず，MASt3R を用いて三次元再構成を行い，三次元空間上で消去対象となる物体を除外した結果を示す。図 4 には，手を消去した後の再構成結果を，入力画像と同じ視点から再レンダリングした画像を示す。図 4 より，消去対象である手の三次元点群が適切に除去されており，その背後に存在する背景が，他の視点から得られた情報に基づいて補完されていることが確認できる。

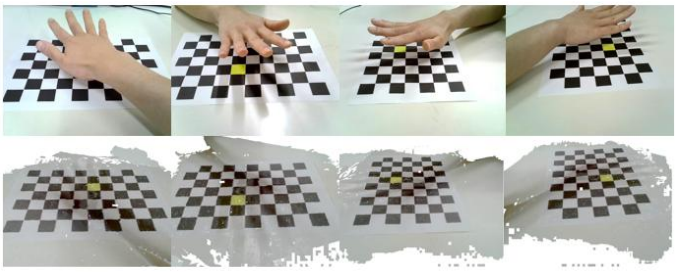

図 4：入力画像(上段)、再構成結果(下段)

4.2. Gemini の画像生成で行った結果

次に、4.1 の結果から二次元画像最適化を適用した最終結果を示す。 図 5 には，提案手法による最終出力(右)と， 既存の拡散モデル Stable Diffusion [2]を用いたインペイントによる結果 (左) を比較して示す。Stable Diffusion によるインペイントでは，さらに ControlNet [3]を用いて他視点画像を入力補助として利用している。拡散モデルによるインペイントでは，消去領域が周囲の文脈と調和するように生成されているものの，背景の構造や形状が実際のシーンと一致しない場合が見られる。特に，背景に規則的な構造や連続性が存在する場合には，生成結果に歪みや不自然な補完が生じる傾向が確認された。一方，提案手法では，三次元再構成に基づいて背景を補完した後に画像最適化を行っているため，背景構造の整合性が保たれている。さらに，Gemini を用いた最終的な画像生成処理により，消去領域周辺のノイズや微小な隙間が抑制され，視覚的に自然な最終結果が得られていることが分かる。

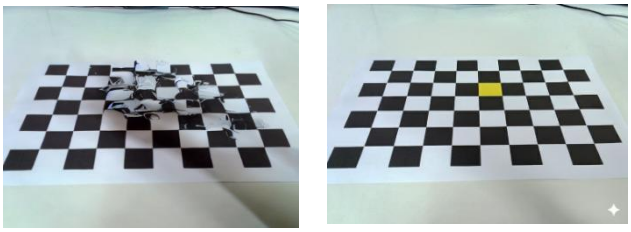

図 5：Stable Diffusion による結果(左)、最終結果(右)

5．　今後の展望

まず，本手法は現時点では静止画像を対象としており，動画への適用は行っていない。そのため，時間方向の連続性を考慮した三次元再構成および物体消去を行うことで，動画全体に対して一貫した消去結果を得られるようにすることが今後の課題である。特に，フレーム間での幾何的一貫性や，消去結果の時間的な安定性をどのように維持するかが重要となる。また，本研究の設定では，消去対象の背後が他の視点画像に写っている場合にのみ，その情報を用いて背景を補完している。一方で，他の視点にも写っていない領域については，三次元的な実情報が存在しないため，本手法のみでは正確な復元が困難である。今後は，三次元再構成に基づく情報と生成モデルによる補完をより密接に統合し，観測されていない領域についても自然かつ整合的に補完できる仕組みの構築を検討する。さらに，本研究の最終結果では，消去対象そのものは高い精度で除去できているものの，床や壁に落ちる影などの間接的な影響までは十分に考慮できていない。影はシーンの照明条件や幾何構造と密接に関係しており，視覚的な自然さに大きく影響する要素である。今後は，照明や影の情報を考慮した三次元表現や，影の推定・除去を組み込むことで，より完全な物体消去を目指す。

参考文献

[1] Vincent Leroy, Yohann Cabon, Jerome Revaud, Grounding Image Matching in 3D with MASt3R

[2] Robin Rombach,Andreas Blattmann, Dominik Lorenz1, High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

[3] Lvmin Zhang, Anyi Rao, Maneesh Agrawala, Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2023