ドローンを用いた 3 次元再構成における テキスト指示型編集技術の開発

菊地佑太

内容梗概

環境の画像から自由視点画像を得る技術は，近年の AI の発達によって急激に発展している．3D Gaussian Splatting（3DGS）はその代表的な例である．ドローンを用いて屋外撮影をする場合の固有の課題が発生するが，それらに対する先行研究もある．例えば，時刻とともに人や物の位置が変化してしまうため，3 次元再構成が困難になるが DroneSplat ではそれに対する対応が行われている．

本研究では映り込んだ不要物体を除去することを考える．提案手法では，テキストで除去する物体を指定することで，自由視点画像から所望の物体を除去することができる．これにより，ドローン画像の 3 次元再構成の活用がしやすくなる．これを実現するために，学習用画像から不要な物体を CLIP 等のマルチモーダル画像分類器で同定し，学習に使用する画像から削除することで，再構成させる自由視点画像の最適化を試みる．

東北大学大学院情報科学研究科　応用情報科学専攻　学位論文, C4IM4010

目次

1　はじめに

1.1　本研究の概要

環境の画像から自由視点画像を得る技術は，近年の AI の発達によって急激に発展している．3D Gaussian Splatting（3DGS）[1] はその代表的な例である．ドローンを用いて屋外撮影をする場合の固有の課題が発生するが，それらに対する先行研究もある．例えば，時刻とともに人や物の位置が変化してしまうため，3 次元再構成が困難になるが（図 1-a），DroneSplat[2] ではそれに対する対応が行われている（図 1-b）．

本研究では映り込んだ不要物体を除去することを考える．提案手法では，テキストで除去する物体を指定することで，自由視点画像から所望の物体を除去することができる．これにより，ドローン画像の 3 次元再構成の活用がしやすくなる．これを実現するために，学習用画像から不要な物体を CLIP[3] 等のマルチモーダル画像分類器で同定し，学習に使用する画像から削除することで（図 2），再構成させる自由視点画像の最適化を試みる．

1.3DGS[1]　2.DroneSplat[2]　3.ravi2024sam2[4]　4.InpaintAnything[5]　5.CLIP[3]　6.LaMa[6]　7.MVinpainter[7]

表1: Simingshan データセットにおける各手法の性能比較

Method	Simingshan		
	PSNR↑	SSIM↑	LPIPS↓
NeRF[8]	19.07	0.417	0.267
3DGS[1]	19.68	0.476	0.254
DroneSplat[2]	22.76	0.759	0.152
Ours	22.35	0.744	0.174

(a) 3DGS[1] で作成
車が動いている場面を表現できない

(b) DroneSplat[2] で生成
動体を無視して表現できる

図1: 動的シーンにおける自由視点画像生成の例

1.2　本研究の構成

本論文は次のように構成される。2章では... 本論文は次のように構成される。2章では...

2　研究背景

2.1　関連研究

3 次元再構成技術

3 次元再構成とは，3 次元空間における物体の形状や位置を再構成する技術である．SfM（Structure-from-Motion：運動復元）[9] はその代表的な例である．SfM は，同一シーンを異なる視点から撮影した複数の画像から，3 次元空間におけるシーンの構造（Structure）と各画像を撮影したカメラの位置・姿勢（Motion）を同時に推定する技術である．処理の流れとしては，まず各画像から特徴点（SIFT や ORB などの特徴記述子）を検出し，異なる画像間で対応する特徴点をマッチングする．次に，対応点の情報を用いてカメラの内部パラメータと外部パラメータ（位置・姿勢）を推定し，三角測量により 3 次元点群を復元する．最後に，バンドル調整（Bundle Adjustment）と呼ばれる最適化手法により，再投影誤差を最小化することで，カメラパラメータと 3 次元構造を同時に精密化する．SfM は，特別な計測機器を必要とせず，一般的なカメラで撮影した画像のみから 3 次元モデルを構築できる点が大きな利点であり，フォトグラメトリやデジタルアーカイブ，地図作成などの幅広い分野で活用されている．

そのほかにも，ニューラルネットワークを用いて三次元空間を表現する手法がある．NeRF[8] は，2020 年に Mildenhall らによって提案された，ニューラルネットワークを用いてシーンを連続的な関数として表現する技術である．同一シーンを異なる視点から撮影した複数の画像とカメラパラメータを用いて学習し，ボリュームレンダリングにより任意の視点から高品質な自由視点画像を生成することができる．NeRF の主な利点は，従来の点群ベースの手法と比較して，三次元空間を連続的な関数として表現することで，細部まで高品質な画像を生成できる点である．一方で，レンダリング時に各ピクセルごとにニューラルネットワークの推論が必要となるため，レンダリング速度が非常に遅く，リアルタイムでの自由視点ナビゲーションが困難であるという課題がある．

このレンダリングの遅さを解消したのが，2023 年に発表されたのが Kerbl らによって提案された 3D Gaussian Splatting (3DGS) [1] である．学習を進めながら生成の品質を高める点では NeRF と同一であるが，三次元空間を 3 次元ガウス分布の集合として表現することが特徴である．各ガウス分布は，位置，共分散行列（形状と向き），不透明度，そして球状調和関数に基づく色情報をパラメータとして持つ．ポイントベースの α（アルファ）ブレンディングと NeRF スタイルのボリュームレンダリングは，本質的に同じ画像生成モデルを共有している．具体的には，ある光線（レイ）に沿った色

C は，以下のボリュームレンダリングの式で与えられる：

$$C = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))c_i \quad \text{ここで} \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \tag{1}$$

ここで，密度 σ，透過率 T，および色 c のサンプルが，間隔 δ_i で光線に沿って取得される．これは次のように書き換えることができる：

$$C = \sum_{i=1}^{N} T_i \alpha_i c_i \tag{2}$$

ただし，$\alpha_i = (1 - \exp(-\sigma_i \delta_i))$ および $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$ である．典型的なニューラルポイントベースのアプローチは，ピクセルに重なる N 個の順序付けられた点をブレンディングすることで，ピクセルの色 C を計算する：

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j) \tag{3}$$

ここで，c_i は各点の色であり，α_i は学習された「点ごとの不透明度」に，共分散 Σ を持つ 2 次元ガウス関数を評価して得られる値を掛け合わせることで求められる．

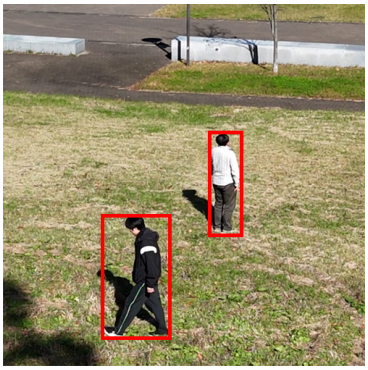
自由視点画像を生成する際，これらの 3D ガウス分布をカメラ平面にスプラッティング（射影）し，ハードウェアが高速に処理できるラスタライズ技術を利用してピクセル単位の色を合成する．これにより，NeRF と比較して画質を維持しつつ，学習時間およびレンダリング速度を大幅に短縮し，リアルタイムでの自由視点ナビゲーションを可能にした [1]．本研究の提案手法も，この高速な 3DGS を基盤技術として採用している．

セマンティックセグメンテーション

セマンティックセグメンテーションは，画像の各ピクセルをなんらか意味を持った領域に分割する技術のことである．一つの画像に対して一つのラベルを付与する画像分類問題や，画像内の物体を認識し，矩形で過酷む物体検出と比べると，セマンティックセグメンテーションは画像中のピクセルに対してラベルを付与することで，より細かい粒度での物体の分割を行うことができる（図 2）．

SAM2[4] はセマンティックセグメンテーションの代表的な例である．SAM2 は，従来のセマンティックセグメンテーション手法が特定の物体クラスに限定されていたり，動画シーケンスへの対応が困難であったりした課題を克服した画期的な手法である．SAM2 の最大の特徴は，大規模データセットで学習された汎用的なセグメンテーションモデルを持ち，事前に学習した物体クラス以外の物体もセグメンテーションできる点である．さらに，画像だけでなく動画シーケンスにも対応し，時系列情報を活用することで，より一貫性のあるセグメンテーション結果を生成することができる．また，ユーザーからの点や矩形，テキストなどの多様な入力形式を受け付けることができる点も，従来手法と比較して大きな進歩であった．

(a) 画像分類
画像全体を見て「人」であると判断している



(b) 物体検出
画像内の物体を矩形で囲んで認識し，
それぞれの物体の位置とクラスを特定している

(c) セマンティックセグメンテーション
画像の各ピクセルに対してラベルを付与することで，
より細かい粒度での物体の分割を行うことができる

図2: コンピュータビジョンタスクの比較

2.2　現状の課題

3　提案手法

3.1　不要物同定アルゴリズム

3.2　拡散モデルによる削除部の補完

4　評価実験と考察

4.1　データセット

4.2　学習

4.3　評価手法

4.4　結果

5　おわりに

おおおおおおおおおおおお

5.1　結論

5.2　今後の課題

参考文献

[1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[2] Jiadong Tang, Yu Gao, Dianyi Yang, Liqi Yan, Yufeng Yue, and Yi Yang. Dronesplat: 3d gaussian splatting for robust 3d reconstruction from in-the-wild drone imagery, 2025.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[4] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[5] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[6] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

[7] Chenjie Cao, Chaohui Yu, Yanwei Fu, Fan Wang, and Xiangyang Xue. Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. *arXiv preprint arXiv:2408.08000*, 2024.

[8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.

[9] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. In *ACM SIGGRAPH*, pages 835–846, 2006.