

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。

目次

1	はじめに	3
1.1	本研究の概要	3
1.2	本研究の構成	4
2	研究背景	5
2.1	関連研究	5
2.2	課題	5
3	提案手法	6
3.1	不要物同定アルゴリズム	6
3.2	拡散モデルによる削除部の補完	6
4	評価実験と考察	7
4.1	データセット	7
4.2	学習	7
4.3	評価手法	7
4.4	結果	7
5	おわりに	8
5.1	結論	8
5.2	今後の課題	8

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

[illegible]

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。NeRF[8] はその代表的な例である。このレンダリング速度の課題を克服するために、2023年に発表されたのが 3D Gaussian Splatting (3DGS) [1] である。3DGS は、シーンを膨大な数の **3次元ガウス分布 (Gaussian)** の集合として表現する。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 (Spherical Harmonics: SH) に基づく色情報をパラメータとして持つ。3DGS の最大の特徴は、レンダリングプロセスにある。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。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。

2.2 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

oooooooooooooooo

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, Dianyí Yang, Liqi Yan, Yufeng Yue, and Yi Yang. Drone-splat: 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. Mvin-painter: 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.