単一画像からの段階的視野拡張と 3 次元点群化とを統合した VR 画像合成手法

A VR Image Synthesis Method Integrating Single-Image 3D Point Cloud Generation and Field-of-View Expansion

的場未奈 [1a)]，戸辺義人 [1b)]，鷲見和彦 [1c)]
Mina MATOBA, Yoshito TOBE, and Kazuhiko SUMI

1) 青山学院大学 （〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1）
a) c5624217@aoyama.jp b) tobe@it.aoyama.ac.jp c) sumi@it.aoyama.ac.jp

概要：VR 用画像合成において，視野角 60° の単一画像から視野角 120° の 3 次元（3D）点群を構築する手法を提案する．画像を左右 30 % ずつ逐次生成し，その都度深度推定して 3D 点群化し，それらを位置合わせして統合することで，広視野 3D 点群を構築する．既存の一括拡張法 VistaDream と比較し，視野角 120° パノラマ画像の拡張領域が実写画像と意味一貫性を保つかを評価した結果，6 シーン平均で SSIM が 32 % 向上し，LPIPS が 12 % 低減した．

キーワード：単一画像三次元再構成，Novel View Synthesis，VR 没入感

1. はじめに

近年の VR (Virtual Reality) では，ロボットやユーザが仮想空間内を移動する際にも違和感のない視覚体験を提供するため，高忠実 3 次元 (3D) シーン生成が不可欠である．従来は生成された 3D モデルにテクスチャを貼付して表現する手法が主流であった．しかし，NeRF (Neural Radiance Fields) [1] や 3DGS (3D Gaussian Splatting) [2] 等の NVS (Novel View Synthesis) [3] の登場により，多視点の実写画像から新規視点画像を直接合成できるようになった．これにより，複雑なモデリングなしに忠実な多視点画像の生成が可能となった．

NeRF は，多数の入力画像を用いて光線上の輝度と密度を推定し，高品質な自由視点レンダリングを実現する．しかし，3 次元空間全体 (シーン) に対して長時間の学習を要し，多視点画像を前提とする点が課題である．これに対して 3DGS は，画像から得た粗点群を，位置・大きさ・色を持つ 3 次元 Gaussian 分布 (粒子) として保持し，粒子を画像平面上に直接描画するため， NeRF と比べレンダリングを大幅に高速化し，短時間の最適化でも高精細な描画を実現する．しかし 3DGS も多視点入力を必要とする．そのため単一画像しか得られない状況では適用が難しい．

この制約を解消するため，VistaDream [4] は拡散モデルと深度推定を組み合わせ，単一画像から高忠実 3D GF (Gaussian Field) を生成する．まず周辺 45 % の拡張画像を生成し (outpainting) [5]，得られた RGB-D (Depth) を粗い 3D GF に変換する．その後，粗い 3D GF を異なる視点からレンダリングして得た複数画像に対し，幾何形状の整合性を高める MCS (Multiview Consistency Sampling) [4] を適用すると同時に，ノイズを除去して 3D GF へ統合し，最終的に高忠実 3D GF を得る．

VistaDream において，拡張画像生成するにあたり，画像全体を要約して自然言語で出力可能な LLaVA (Large Language-and-Vision Assistant) [6] で，入力画像からテキストプロンプトを作成する．その後，作成されたテキストから入力画像の周辺コンテキストを推定して outpainting し，拡張画像を生成する．

しかし，VistaDream で生成された 3D GF を VR 上で確認すると，45 % の outpainting では， 視野角 (FoV) 90° まで拡張しているが，VR HMD (Head Mounted Display) の水平視野を (FoV ≈ 120°) を覆えず，画面端に未生成領域が残ることが報告されている [5]．さらに同入力画像から LLaVA が生成したテキストで FoV 120° の高忠実 3D GF を生成すると，生成領域に意味的一貫性のない箇所（semantic drift）が散見された．

そこで我々は，入力画像の左右 30 % ずつ段階的に拡張し，逐次 LLaVA を用いて画像からテキストを生成し，そのテキストを基に次の拡張画像を生成した．各画像を 3D GF に変換し統合することで，画面端部の高解像度サンプリングを確保しつつ，各生成ステップのコンテキストを維持して semantic drift を抑制した 3D GF の実現を目指す．

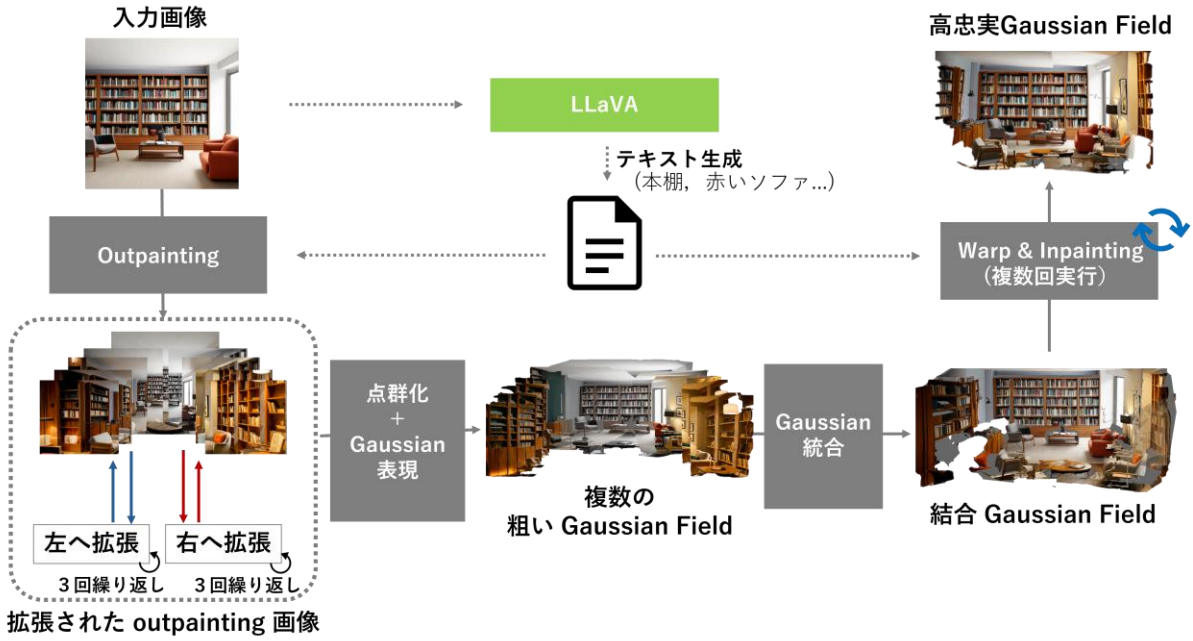

図 1：提案手法のシステム図

以下，本論文第 2 章では提案手法の詳細を説明する．第 3 章では定性・定量評価の結果を示し，第 4 章で結論を述べる．

2. 提案手法

本章では，単一入力画像を FoV 120° まで拡張し，高忠実 3D GF を生成する手法を述べる．

2.1 Outpainting

本提案手法のシステム全体図を図 1 に示す．入力画像に対し，LLaVA を用いてテキストを生成する．生成されたテキストを入力として画像を逐次 outpainting することで，入力画像と意味的一貫性を保った拡張画像を得られる．本提案では，各ステップの拡張率を 30 % に抑えることで，入力画像の意味的コンテキストを十分に保つことを前提として設計する．

ここで，入力画像の FoV を FoV_{in}，最終的に求めたい FoV を FoV_{out}とする．さらに，入力画像を I_0，画像の高さと幅をそれぞれ H と W，左右 30 % 拡張する outpaint 演算を $OP_{0.3}(\circ)$ とし，拡張後の画像から横幅 W を保持したまま $H \times W$ 領域を切り出す crop 演算を，左側を C_L，右側を C_R と定義する．

左側および右側に outpainting される画像列を，各々，I_{2k-1}，I_{2k} $(k = 1,2,3,\dots)$ とする．左右各々 N 個に対する I_n は，以下の擬似コードに示される要領で求まる．

Initialization: $J_0 \leftarrow OP(I_0)$

$I_1 \leftarrow C_L(J_0)$

$I_2 \leftarrow C_R(J_0)$

$for\ k: 1 \sim N$:

$J_{2k-1} \leftarrow OP(I_{2k-1})$

$J_{2k} \leftarrow OP(I_{2k})$

$I_{2k+1} \leftarrow C_L(J_{2k-1})$

$I_{2k+2} \leftarrow C_R(J_{2k})$

FoV_{out} を 120° にするために，$N = 4$ とし，その結果，I_k $(k = 0,\dots,8)$ が得られる．本研究では，個々の入力画像 I_k を独立に 3DGS パイプラインへ投入し，内部で生成される粗い 3D GF を $S_k = GS(I_k)$ として抽出する．各 S_k は単眼深度推定に起因する位置誤差を含むため， 2.2 で述べる coarse-fine ICP (Iterative Closest Point) を使用した GaussReg [7] により逐次位置合わせし，最終的な高忠実 3D GF を得る．

2.2 3D GF の逐次統合

拡張画像列 I_k $(k = 0,\dots,8)$ から得た 3D GF S_k $(k = 0,\dots,8)$を生成順に入力とする．

最初に Geo Transformer [8] で得られた特徴で粗姿勢を推定し，点対平面 ICP [9] で 並進 1 mm 未満・回転 0.05° 未満に収束させる．各ステップの点群は重み付き平均で統合し，半径融合で冗長点を削減する．最後に pose-graph 最適化 [10] でループ誤差を抑え，累積誤差を最小化した統合 3D GF を得る [7].

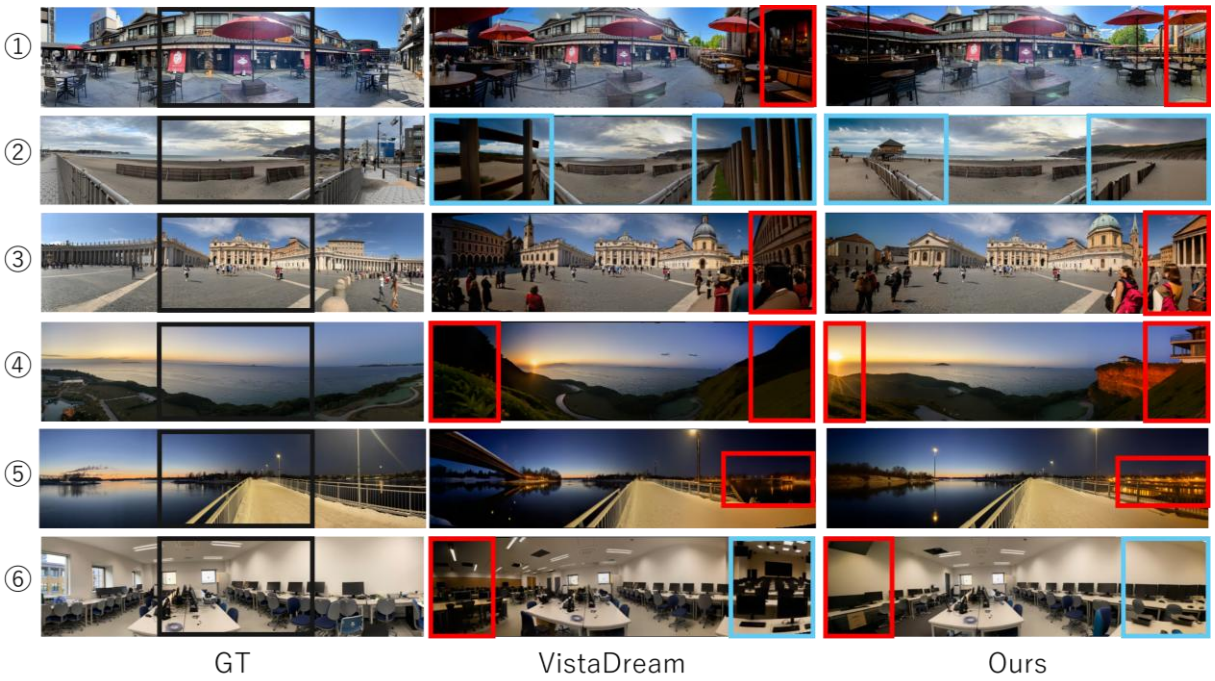

図 2:VistaDream と 提案手法により生成された結果（左＝GT，中央＝VistaDream，右＝Ours）．左：GT パノラマ画像．画像中央部を入力画像としている（黒枠）．中央：VistaDream は一括 45 % outpainting のため端部に暗転（赤枠）や不自然物（青枠）が残る．右：提案手法は 30 %×8 ステップ outpainting＋逐次整合により，視野 120° でも端部破綻がほぼ見られない．

表 1: 図 2 の 6 シーンに対する定量評価結果

	SSIM↑		LPIPS↓	
	VistaDream	Ours	VistaDream	Ours
①	0.2045	0.2289	0.6728	0.6335
②	0.1482	0.2756	0.6597	0.5175
③	0.2389	0.2744	0.5658	0.5213
④	0.4536	0.5806	0.6047	0.5232
⑤	0.3788	0.4985	0.5370	0.4265
⑥	0.2502	0.3474	0.5743	0.5301

3. 実験

本章では，本研究の評価方法とその結果を示す．我々が撮影した 6 種類のパノラマ画像 (①〜⑥) を用いて，定性的評価では視覚的一貫性を検証し，定量的評価では画質指標と点群指標の両面から VistaDream との比較を行う．

3.1 定性的評価

本節では，拡張領域を含む 120° パノラマ画像について定性的に比較する．図 2 に，実写画像 (GT)，既存手法（VistaDream），提案手法 (Ours) の比較結果を示す．

VistaDream では，

- ②，⑥ 画像について，新たな空間や生成物 (青枠) が生成されていることが確認できる．
- 拡張領域（赤枠）が一段暗く，照度不連続が生じる．

など，semantic drift と幾何的不整合が同時に観察される．

一方，提案手法では

- 拡張後も追加の部屋は生成されない．
- 輝度・色調が中心視野と一致し，暗転が極めて稀である．

ことを確認できる．すなわち，段階的 outpainting により，semantic drift を抑制しながら $FoV_{out} = 120°$ を実現可能であることが確認できた．

3.2 定量的評価

拡張領域のみをマスクしたうえで，構造的類似度を測る SSIM (Structural SIMilarity) [11] と，近く距離を測る LPIPS (Learned Prrceptual Image Patch Similarity) [12] を使用した．SSIM は 1.0 が完全一致，LPIPS は 0.0 が最良．評価は拡張領域のみを自動マスクし算出した．

表 1 に，SSIM と LPIPS によるシーン別数値を示す．6 シーン平均で，VistaDream は SSIM 0.279／LPIPS 0.602，Ours は 0.368／0.527 を達成した．SSIM は +0.089（＋32 %）向上し，LPIPS は −0.075 (−12 %) 低減した．

片側 t 検定（$p < 0.05$）により両指標で統計的有意差が確認された．とくに 図 1 の④シーンでは SSIM が 0.5806 と最高値を示し，図 1 の⑤でも LPIPS を 0.4265 まで削減している．

低い絶対値（SSIM < 0.6，LPIPS > 0.4）は，拡張領域が GT に存在しないためであり，定性的比較（図 1）でも確認できるように，提案手法は視野端での semantic drift を

大幅に抑制している．

4. 結論

本研究は段階的 outpainting と逐次 3DGS の整合により，semantic-drift を抑えつつ視野角 120° パノラマ画像を生成し，VistaDream 比で SSIM＋32 %，LPIPS-12 %の有意な改善を示した．現状の GaussReg は剛体（回転・並進）で全体を一括整合しているため，今後は局所伸縮を許容して，形状歪みをさらに低減し，複雑シーンへの適用を目指す．

参考文献

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Proc. Eur. Conf. Comput. Vis. (ECCV 2020)， Lecture Notes in Computer Science, Vol. 12346, pp. 405–421 (2020)

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis: 3D Gaussian Splatting for Real-Time Radiance Field Rendering, ACM Trans. Graph. 42 (4), Article 139, 1‑14 (2023）

[3] R. Tucker and N. Snavely: Single-View View Synthesis with Multiplane Images, Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2020), pp. 551–560 (2020)

[4] H. Wang, Y. Liu, Z. Liu, W. Wang, Z. Dong, and B. Yang: VistaDream: Sampling Multiview-Consistent Images for Single-View Scene Reconstruction, arXiv:2410.16892 (2024)

[5] H. Wang et al. “VistaDream: Sampling Multiview-Consistent Images for Single-View Scene Reconstruction,” GitHub repository, WHU-USI3DV/VistaDream, commit 9a743a9 (Oct. 23, 2024). Available: https://github.com/WHU-USI3DV/VistaDream (accessed Jan. 15, 2025)

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee: LLaVA: Large Language and Vision Assistant via Visual Instruction Tuning, NeurIPS (NeurIPS 2023, Oral), arXiv:2304.08485 (2023)

[7] J. Chang, Y. Xu, Y. Li, Y. Chen, and X. Han: GaussReg: Fast 3D Registration with Gaussian Splatting, Proc. Eur. Conf. Comput. Vis. (ECCV 2024), 407–423 (2024)

[8] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu: GeoTransformer: Unifying Alignment and Correspondence for Point Cloud Registration, IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 9806–9821 (2023)

[9] P. J. Besl and N. D. McKay: A Method for Registration of 3-D Shapes, IEEE Trans. on Pattern Anal. Mach. Intell., Vol. 14, No. 2, pp. 239–256, 1992.

[10] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard: g2o: A General Framework for Graph Optimization, Proc. the IEEE Int. Conf. on Robotics and Automation (ICRA 2011), pp. 3607-3613 (2011)

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli: Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Trans. Image Process., Vol. 13, No. 4, pp. 600-612 (2004)

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, Proc. the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 586-595 (2018)