레터논문 (Letter Paper)
방송공학회논문지 제30권 제5호, 2025년 9월 (JBE Vol.30, No.5, September 2025)
https://doi.org/10.5909/JBE.2025.30.5.827
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

3D Gaussian Splatting의 초기 포인트 생성을 위한 VGGT와 SfM의 비교

구 래 건[a)], 박 준 형[a)] 라 현 재[b)], 류 은 석[a)‡]

Comparison of VGGT and SfM in Generating Initial Points for 3D Gaussian Splatting

Reagan Koo[a)], Jun-Hyeong Park[a)], Hyunjae Ra[b)], and Eun-Seok Ryu[a)‡]

요 약

3D Gaussian Splatting (3DGS)은 최근 실시간 고품질 렌더링이 가능한 Novel View Synthesis 기술로 주목받고 있다. 3DGS는 초기화 단계로서 주로 COLMAP을 이용한 Structure from Motion (SfM) 기반의 반복적 최적화 방식으로 추정한 희소한 포인트 클라우드로 학습을 진행한다. 한편 최근 등장한 딥러닝 기반 모델 Visual Geometry Grounded Transformer (VGGT)는 이와 달리 feed-forward pass 를 통해 보다 밀집된 포인트 클라우드를 빠르게 생성한다. 본 논문에서는 3DGS 학습을 위해 VGGT와 COLMAP을 사용했을 때의 결과를 비교 분석하였다. 실험 결과, 최종 성능은 유사하였으나 수렴 양상은 초기화 방법과 데이터셋 특성에 따라 차이를 보였다. 복잡하고 넓은 야외 장면에서는 VGGT가 수렴 속도를 단축시키며 안정적인 성능을 보였고, 구조적 정합성이 중요한 실내 장면에서는 COLMAP 초기화가 더 우수한 결과를 나타냈다.

Abstract

3D Gaussian Splatting (3DGS) has recently attracted attention as a novel view synthesis method for real-time high-quality rendering. Typically, 3DGS is initialized with a sparse point cloud reconstructed by COLMAP-based Structure-from-Motion (SfM), whereas the deep learning-based Visual Geometry Grounded Transformer (VGGT) quickly generates denser point clouds through a feed-forward pass. This paper presents a comparison of VGGT and COLMAP for 3DGS initialization. Experimental results show that while the final performance is similar, convergence patterns differ: VGGT accelerates convergence and provides stable performance in complex outdoor scenes, whereas COLMAP initialization yields better results in structured indoor environments.

Keyword : 3D Gaussian Splatting, Structure-from-Motion, VGGT, Novel View Synthesis, Point Cloud

a) 성균관대학교 실감미디어공학과(Department of Immersive Media Engineering, Sungkyunkwan University)
b) 성균관대학교 컴퓨터교육과(Department of Computer Education, Sungkyunkwan University)
‡ Corresponding Author : 류은석(Eun-Seok Ryu)
E-mail: esryu@skku.edu
Tel: +82-2-760-0677
ORCID: https://orcid.org/0000-0003-4894-6105

※ This work was supported by MCST and KOCCA under the Culture Technology R&D Program (No. RS-2023-00223812), and by NRF grant funded by MSIT, Korea (No. RS-2024-00457605)
· Manuscript August 1, 2025; Revised September 8, 2025; Accepted September 8, 2025.

Ⅰ. 서 론

3D 가우시안 스플래팅 (3DGS)[1]은 3차원 공간을 명시적으로 표현하며 높은 품질의 실시간 렌더링을 실현하는 방법으로 주목받고 있다. 3DGS는 장면을 다수의 3차원 가우시안으로 표현하고, 위치·크기·회전·불투명도·구면 조화 계수 등을 최적화함으로써 연속된 시점에서의 이미지를 효율적으로 재구성한다. 이 과정에서 사용되는 초기 포인트 클라우드는 학습 안정성, 손실 수렴 속도, 나아가 최종 표현의 품질까지 좌우하는 핵심 요소로 기능한다. 전통적으로 초기 포인트는 SfM[2] 계열 방법에 의존해 왔다. 그중에서도 COLMAP은 특징점 검출과 정합, 카메라 파라미터 추정, 번들 조정 (Bundle Adjustment, BA)에 이르는 반복적 최적화 절차를 통해 기하적으로 일관된 희소 포인트 클라우드를 산출한다. 이러한 기하 기반 접근은 안정성과 해석 가능성 측면에서 강점을 보이지만, 텍스처가 빈약하거나 반복 패턴이 많은 영역에서 포인트가 충분히 확보되지 않는 한계가 있다. 이에 Visual Geometry Grounded Transformer (VGGT)[3]는 2025 CVPR에서 Best Paper Award를 받으며 그 대안으로서 제안되고 있다. VGGT는 대규모 데이터 학습을 통해 단일 feed-forward pass만으로 이미지 셋으로부터 3차원 구조를 직접 예측하는 방식으로 포인트 클라우드를 생성한다. 본 연구는 이러한 배경에서 3DGS 파이프라인의 초기 포인트 선택 전략이 후속 학습 역학과 최종 재구성 품질에 미치는 영향을 비교하고자 한다. 이를 위해 동일한 장면 세트와 학습 하이퍼파라미터를 고정한 상태에서, 전통적 SfM 기반인 COLMAP과 딥러닝 기반인 VGGT로 각각 생성한 초기 포인트 클라우드를 3DGS 학습에 투입하고 전 과정을 정밀하게 비교하였다. 이를 통해 3DGS 기반 실시간 렌더링·재구성 시스템에서 장면 특성에 따른 초기화 전략 선택 가이드라인을 제공하며, 향후 대규모 데이터셋이나 제한된 자원 환경에서의 최적 파이프라인 설계에 실질적 근거를 제시한다.

Ⅱ. 관련 연구

두 가지 초기 포인트 클라우드 생성 기술 간의 구조적 차이점을 대조하고, 이들에 의존하는 후속 3DGS 프로세스를 설명한다.

1. 기존 방식의 3DGS 학습

COLMAP으로 대표되는 점진적 재구성 기반 SfM은 명시적 기하학 제약과 반복 최적화를 통해 신뢰도를 높이는 다단계 프로세스다. 먼저 Scale-invariant feature transform (SIFT)[4] 등 feature descriptor로 각 이미지의 지역 특징점을 추출해 희소 특징 집합을 얻고, 이미지 쌍 간 매칭 후 오차를 제거한다. 이어 안정적 삼각측량이 가능한 초기 이미지 쌍으로 3D 모델을 시작해, 이미지를 추가하며 포즈 추정, 포인트 삼각측량을 반복하며 BA로 전역 최적화를 수행한다. 이 때 BA는 최대 $O(n^3)$ 복잡도를 갖기에 대규모 장면에서 병목이 되며, 텍스처 부족·반복 패턴 영역 실패와 희소 포인트 클라우드라는 구조적 한계를 가진다. 이후 SfM이 만든 희소 포인트는 3DGS 초기 가우시안으로 확장된다. 이후 3DGS 학습 과정에서 differentiable rasterization으로 렌더링한 결과와 원본 이미지 간 손실을 역전파해 모든 파라미터를 end-to-end로 최적화하며, 초반엔 희소성을 보완하기 위해 화면 그라디언트/재투영 오차 기반 densification과 기여도 낮은 가우시안 pruning을 수행한다. 초기 포인트의 밀도·분포가 수렴 속도와 안정성을 좌우한다는 점에서 희소 초기화는 초반 수렴 지연과 특정 영역 보완 한계를 남긴다.

2. VGGT

VGGT는 방대한 데이터셋으로부터 3D 장면의 기하학적 원리를 사전 학습한 12억 개 파라미터의 대규모 순방향 신경망으로, 기존 다단계 파이프라인을 단일 모델로 대체했다. VGGT는 입력으로 여러 장의 이미지 집합을 받아 각 이미지를 패치로 분할한 뒤, 자기지도 학습 기반의 시각 표현 학습 모델인 DINO (Distillation with No Labels) encoder를 통해 시각 임베딩으로 변환한다. 이후 크로스 어텐션 트랜스포머가 프레임 단위의 로컬 셀프 어텐션과 세트 레벨 전역 셀프 어텐션을 교차 수행함으로써, 각 이미지 내부의 세부 정보와 이미지 집합 전체의 전역 대응 관계를 동시에 처리한다. 최종 출력은 별도의 prediction head로 전달되어 모든 카메라 파라미터, 픽셀 단위 깊이 맵, 그리고

각 픽셀의 3D 좌표를 담은 point map을 직접 추론한다. COLMAP이 희소 특징점을 삼각측량해 포인트를 얻는 것과 달리 VGGT는 모든 픽셀에 대한 3D 좌표를 직접 예측하므로 보다 조밀한 포인트 클라우드를 생성한다. 이 구조 덕분에 반복 최적화를 생략하고 forward pass만으로 수 초 내 결과를 산출하며, 경우에 따라 최적화 기반 대안보다 우수한 성능을 보이기도 한다. 다만 최대 정확도를 위해 선택적으로 BA 후처리를 적용할 수 있다.

Ⅲ. 실험 및 실험결과

본 연구는 초기 포인트 클라우드 생성 방식만을 달리한 두 조건, COLMAP과 VGGT가 3DGS 학습에 미치는 차이를 정량적으로 비교하기 위해 Nerf360[4] 데이터셋의 bicycle, bonsai, garden을 일부 샘플링한 데이터와 Bartender, VRroom1D, VRroom2D[5]를 사용해 RTX 5090 GPU 환경에서 학습을 수행하였다. COLMAP 조건은 특징점 기반 점진적 재구성 파이프라인을 통해 얻은 희소 포인트를, VGGT 조건에서는 사전학습된 모델의 feed forward 추론으로 얻은 카메라 파라미터와 포인트를 초기값으로 사용한 뒤, 추가적인 BA로 기하적 일관성을 정제하였다.

동일한 3DGS 학습 파이프라인에서 COLMAP 초기화와 VGGT 초기화를 적용하여 여섯 개의 장면에 대해 비교 실험을 수행하였다. 표 1 및 그림 1에서 VGGT 초기화는 전

표 1. 각 데이터셋별 학습 결과
Table 1. Training Results for Each Dataset

Dataset	Initial Pts (C/V)	Elapsed Time (s)	VRAM Usage (GB)	100 iter PSNR (dB)	10k iter PSNR (dB)	10k iter SSIM	99% PSNR (min)
bicycle	2527 / **24266**	**4** / 92	**2.13** / 21.1	11.42 / **12.81**	**14.07** / 13.98	**0.48** / 0.47	10.8 / **1.5**
bonsai	8077 / **32504**	**29** / 88	**1.01** / 19.8	11.01 / **11.77**	13.31 / **17.77**	0.51 / **0.77**	2.2 / **1.3**
garden	19271 / **41055**	**27** / 91	**2.12** / 19.9	13.86 / **14.37**	15.00 / **15.47**	**0.59** / 0.58	2.9 / **2.3**
Bartender	11234 / **28608**	**15** / 53	**1.32** / 15.6	23.42 / **23.71**	**33.85** / 33.15	**0.93** / 0.92	2.3 / **1.6**
VRroom1D	7866 / **14208**	**17** / 66	**1.41** / 17.1	**17.47** / 14.65	27.34 / **27.70**	**0.93** / 0.92	**1.5** / 3.2
VRroom2D	5029 / **28538**	**13** / 63	**1.70** / 20.5	**17.32** / 16.68	**26.70** / 24.91	**0.92** / 0.89	1.6 / **0.8**

표 1은 COLMAP(C)과 VGGT(V) 초기화가 3DGS 학습에 미친 영향을 데이터셋별로 비교한 결과를 보여준다. Initial Pts는 학습 시작 시 투입된 초기 포인트 수, Elapsed time은 초기화 수행 시간, VRAM usage는 데이터 셋 처리 시 사용된 평균 vram 사용량, 1k iter PSNR은 1000 iteration 시점의 테스트 PSNR, 10k iter PSNR은 10,000 iteration 종료 시 최종 PSNR. 10k iter SSIM은 종료 시 최종 SSIM, 99% PSNR(min)은 최종 PSNR의 99% 수준에 처음 도달하기까지 소요된 시간

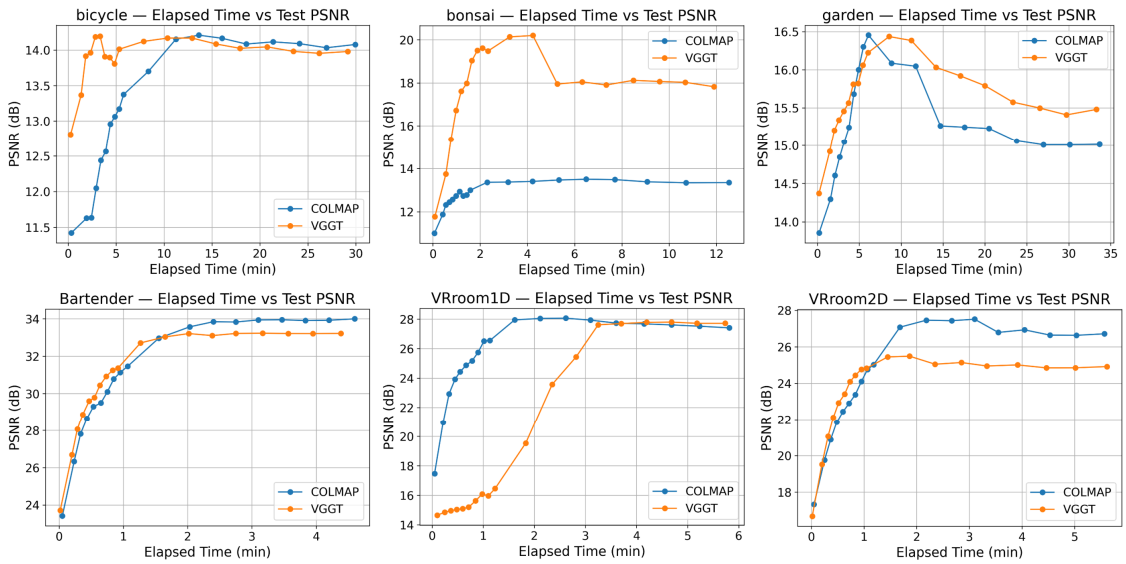

그림 1. 3DGS 학습 PSNR의 시간에 따른 변화 (COLMAP vs VGGT)
Fig. 1. 3DGS Training PSNR vs. Time (COLMAP vs VGGT)

반적으로 더 많은 초기 포인트를 제공함으로써 학습 초기에 빠른 PSNR 향상을 유도하였으며, 대규모 및 복잡한 장면에서 안정적인 수렴 특성을 보였다. 또한 대부분의 경우 99% PSNR 도달 시간이 단축되어 효율적인 학습 진행이 가능하였다. 반면 COLMAP 초기화는 상대적으로 적은 초기 포인트를 기반으로 하여 학습에 필요한 VRAM 요구량이 낮고, 최종 SSIM에서 우위를 보였다. 이러한 결과는 VGGT 초기화가 대규모 장면 처리 및 빠른 수렴에 효과적임을 시사하며, COLMAP 초기화는 제한된 자원 환경이나 구조적 정합성이 강조되는 경우에 유용할 수 있음을 보여준다.

Ⅳ. 결 론

본 연구는 COLMAP과 VGGT 두 가지 초기화 방식을 동일한 3DGS 학습 파이프라인에 적용하여, 학습 효율성과 결과 품질을 비교하였다. 실험 결과, VGGT는 복잡하고 넓은 야외 장면에서 초기 포인트의 양과 분포를 통해 학습 초기를 빠르게 안정화시키고, 최종 품질 역시 유지하거나 개선하는 성과를 보였다. 특히 촬영 뷰가 희소한 야외 환경에서는 COLMAP의 삼각측량 기반 방식이 기하 정보를 충분히 복원하지 못해 카메라 추정이 불안정해졌고, 이로 인해 학습 성능이 크게 저하되었다. 이에 비해 VGGT는 모든 카메라를 안정적으로 추정하고 충분한 초기 포인트를 확보하여 동일한 조건에서도 우수한 성능을 나타냈다. 한편, 경계가 명확하고 구조가 세밀한 실내 장면에서는 COLMAP 초기화가 더 효율적이었다. 실내 환경에서는 밀집된 뷰 분포로 인해 삼각측량 조건이 잘 충족되며, COLMAP의 기하 기반 구조 추정이 효과적으로 작동했기 때문이다. 종합하면, 복잡하고 넓은 장면에서 빠른 반복 학습이 필요한 경우에는 VGGT, 반대로 기하학적 정합성이 중요하고 메모리 제약이 있는 균질한 실내 장면에는 COLMAP 초기화를 사용하는 것이 합리적이다. 향후에는 보다 다양한 시퀀스로 실험을 진행하며 장면 특성에 따라 두 방식을 하이브리드로 적용하거나, VGGT의 초기화 방식에 적합한 3DGS의 densification 전략을 개선하는 연구가 요구된다.

참 고 문 헌 (References)

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3D Gaussian Splatting for Real-Time Radiance Field Rendering,” *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, article no. 139, pp. 1-14, Jul. 2023.
doi: https://doi.org/10.1145/3592433

[2] J. L. Schönberger and J.-M. Frahm, “Structure-from-Motion Revisited,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 4104-4113, Jun. 2016.
doi: https://doi.org/10.1109/CVPR.2016.445

[3] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, “VGGT: Visual Geometry Grounded Transformer,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, pp. 5294-5306, Jun. 2025.
doi: https://doi.org/10.1109/CVPR52734.2025.00499

[4] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, “Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, pp. 5460-5469, Jun. 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00539

[5] J. Choi, Y. Ryu, Y. Choi, J.-B. Jeong, J.-H. Park, I. Yang, and E.-S. Ryu, “[INVR]EE2.1-Related: Report with New Natural INVR Video Contents: SKKU_VRroom,” document MPEG2023/m64721, ISO/IEC JTC1/SC29/WG4, 144th MPEG Meeting, Oct. 2023.