특집논문 (Special Paper)
방송공학회논문지 제30권 제6호, 2025년 11월 (JBE Vol.30, No.6, November 2025)
https://doi.org/10.5909/JBE.2025.30.6.927
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

NeRF 계열과 3DGS 계열 모델의 성능 비교

우 성 현[a)], 서 정 일[a)‡]

Comparative Study on the Performance of NeRF and 3DGS based models

SeongHyun Woo[a)] and Jeongil Seo[a)‡]

요 약

본 논문에서는 하나의 객체를 다양한 구도에서 촬영한 이미지들과 해당 이미지들의 카메라 위치 정보를 활용하여 실제로 촬영하지 않은 새로운 시점을 생성하는 NeRF와 3DGS 및 그 후속 모델들의 성능을 비교·분석한다. 평가 지표로는 시각 품질 지표인 PSNR, SSIM, LPIPS와 학습 효율성을 판단하기 위한 지표로 Training Time, Training Peak Memory를 사용하여 각 모델의 성능을 종합적으로 평가한다. 분석 결과 NeRF 계열 모델은 메모리 사용량 측면에서는 효율적이었으나 학습 시간이 길고 재구성 품질이 낮은 경향을 나타냈다. 반면 3DGS 계열 모델은 학습 속도가 빠르고 재구성 품질이 높아 효율성 및 실용성 측면에서 우수한 성능을 보였으나, 학습 시 메모리 사용량이 크게 증가하는 한계가 있었다. 이러한 결과는 메모리 사용량 증가에도 불구하고 3D 장면 재구성 분야에서는 시각 품질이 더 중요한 요소이기 때문에 3DGS 계열 모델이 NeRF 계열 모델보다 더 실용적인 대안이 될 수 있음을 시사한다.

Abstract

This paper compares and analyzes the performance of NeRF, 3DGS, and their subsequent models for generating novel views of a scene from multi-view images of a single object. The evaluation includes widely used image quality metrics such as PSNR, SSIM, and LPIPS, as well as training efficiency indicators including training time and training peak memory consumption. Through this comprehensive evaluation, the overall performance of each model is assessed. The results show that NeRF-based models are efficient in terms of memory usage but require long training time and tend to deliver lower reconstruction quality. In contrast, 3DGS-based models demonstrate faster training convergence and higher reconstruction fidelity, achieving superior performance in terms of efficiency and practicality, though at the cost of significantly increased peak memory consumption during training. This result suggests that, despite the increase in memory usage, 3DGS-based models can serve as a more practical alternative to NeRF-based models in the field of 3D scene reconstruction, where visual quality is a more critical factor.

Keyword : Neural Radiance Fields, 3D Gaussian Splatting, Novel View Synthesis, 3D Vision, 3D Reconstruction

Ⅰ. 서 론

최근 디지털 콘텐츠는 AR/VR과 같은 몰입형 콘텐츠를 중심으로 단순한 고화질을 넘어 단방향의 시선에 머무르지 않고 장면을 다양한 시점에서 바라보려는 흐름으로 확장되고 있다. 그러나 여전히 현재 대부분의 영상 콘텐츠는 2D 영상 기술에 의존한다. 2D 영상 기술은 대표적으로 일반 비디오 촬영, 파노라마 사진, 그리고 스테레오 비전과 같은 방식들이 있지만, 이들은 기본적으로 제한된 시점에서만 장면을 기록한다. 이로 인해 사용자는 특정 시점에서만 화면을 바라볼 수밖에 없으며 촬영되지 않은 각도에서의 시각 정보를 얻을 수 없다. 이러한 제약은 장면을 자유롭게 탐색하거나 재구성하는 데 큰 한계로 작용한다. 이러한 한계를 극복하기 위해 3D 장면 재구성 기술이 주목받고 있다. 3D 기술은 정해진 시점 외의 새로운 시점을 생성하고 장면을 재구성함으로써 사용자가 원하는 시점에서 자유롭게 장면을 감상하거나 새로운 뷰를 생성할 수 있도록 한다. 그중 특히 주목받고 있는 기술은 NVS (Novel View Synthesis)이다[1]. NVS는 다중 시점에서 촬영된 이미지와 각 이미지에 대응되는 카메라의 위치 정보를 바탕으로 실제로 촬영되지 않은 새로운 시점의 영상을 생성하는 방식이다. 이를 통해 유연한 시각적 경험을 제공할 수 있으며 몰입감 있는 3D 콘텐츠 구현이 가능하다.

이처럼 NVS에 대한 관심이 커지는 가운데, 이를 혁신적으로 구현한 기술인 NeRF (Neural Radiance Fields)가 등장하였다[2]. NeRF는 신경망을 기반으로 장면을 연속적으로 표현함으로써 새로운 시점을 합성할 수 있는 기술로 높은 품질을 제공하지만 학습 속도가 매우 느리다. 또한 고주파 성분이 충분히 샘플링되지 못해 저주파 패턴으로 왜곡되는 aliasing 문제에 취약하다는 한계가 있다. 이러한 한계를 개선하기 위한 다양한 후속 모델 연구가 활발히 이루어졌으며 대표적으로 cone tracing을 도입하여 aliasing 문제를 완화한 Mip-NeRF와 hash encoding을 활용해 학습 속도를 크게 향상시킨 Instant-NGP가 있다[3][4]. 그러나 이와 같은 신경망 기반 방식은 근본적으로 연산 비용이 크고 최적화 속도에 한계가 있다. 이후 제안된 3DGS (3D Gaussian Splatting)는 장면을 3차원 Gaussian 분포 집합으로 직접 표현하여 빠른 학습과 실시간 렌더링을 가능하게 하였다[5]. 그러나 Gaussian 수가 증가할수록 메모리 사용량이 크게 늘어나고 초기 포인트 클라우드 품질에 따라 성능이 좌우되는 한계가 있다. 이를 보완하기 위해 3DGS 계열의 후속 연구가 제안되었으며 대표적으로 multi-scale 표현을 도입하여 aliasing을 억제한 Mip-Splatting과 표면 정렬 방식을 적용해 기하학적 정확도를 높인 2DGS (2D Gaussian Splatting)가 있다[6][7]. NeRF와 3DGS는 모두 NVS를 구현하는 핵심 기법이지만 장면을 구성하는 방식, 결과 품질, 처리 속도 등 여러 측면에서 뚜렷한 차이를 보인다.

본 연구에서는 NeRF와 3DGS 및 그 후속 모델들의 장면 표현 방식의 구조적 차이가 Novel View Synthesis의 성능과 시각 품질에 미치는 영향을 분석한다. 이를 위해 동일한 조건에서 두 계열 모델을 학습시킨 뒤, PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), LPIPS (Learned Perceptual Image Patch Similarity)의 시각 품질 지표와 Training Time, Training Peak Memory를 지표로 학습 효율성을 평가하고 정성적 비교를 종합적으로 수행한다[8][9][10].

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 NeRF와 3DGS의 기본 개념을 정리하고, 각 기술의 한계 및 이를 보완하기 위해 제안된 후속 모델들의 발전 흐름과 특징을 체계적으로 고찰한다. 제Ⅲ장에서는 본 연구에서 채택한 실험 방법론을 상세히 기술하고 실험 환경, 구현 세부 사항, 사용된 데이터셋과 평가 지표를 정리한다. 제Ⅳ장에서는 NeRF와 3DGS 및 그 후속 모델들에 대한 정량적 결과와 정성적

a) 동아대학교 컴퓨터공학과(Dept. of Computer Engineering, Donga-A University)
‡ Corresponding Author : 서정일(Jeongil Seo)
E-mail: jeongilseo@dau.ac.kr
Tel: +82-51-200-7796
ORCID: https://orcid.org/0000-0001-5131-0939
※ 이 논문의 결과 중 일부는 한국방송·미디어공학회 2025년 하계학술대회에서 발표한 바 있음
※This work was partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.2023-0-00076, National Program of Excellence in Software(Dong-A University)) and National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT)(RS-2023-00273349_3)
· Manuscript September 15, 2025; Revised October 14, 2025; Accepted October 14, 2025.

결과를 제시한다. 이를 기반으로 모델별 성능, 효율성, 시각적 품질을 종합적으로 분석하고 나아가 결론에서 3D 장면 재구성 분야에서의 확장 가능성과 향후 연구의 개선 방향을 제안한다.

Ⅱ. 비교 모델 분석

본 절에서는 NeRF와 그 후속 모델 Mip-NeRF, Instant-NGP, 그리고 3DGS와 그 후속 모델 Mip-Splatting, 2DGS의 구조와 특성을 비교한다. 이후 이 모델 간의 차이점과 한계를 비교하여 정리한다.

1. NeRF based model

2020년에 발표된 NeRF는 Novel View Synthesis 연구의 큰 전환점을 마련하였다. NeRF는 공간 좌표와 관측 방향을 입력으로 하여 각 지점의 색상과 밀도를 예측하고, 이를 통해 새로운 영상을 렌더링하는 신경망 기반 모델이다.

1.1 Vanilla NeRF

그림 1과 같이 NeRF는 장면 내의 한 점을 표현할 때 해당 위치 (x, y, z)와 카메라 방향 (θ, ϕ)을 입력으로 받아 해당 위치의 색상 (R, G, B)과 밀도 σ를 출력하는 신경망 F_{Θ}를 학습한다. 이처럼 NeRF는 각 ray를 따라 다수의 샘플 지점을 추출하고 그 결과를 누적하는 volume rendering 기법을 통해 이미지를 생성한다. 이러한 방식은 입체감과 사실감이 뛰어난 이미지 생성이 가능하며 자유로운 장면 표현이 가능하다.

그러나 NeRF는 하나의 픽셀을 생성하기 위해 카메라에서 발사된 ray마다 수십에서 수백 개의 샘플 지점을 통과하며 각 지점마다 MLP(Multi Layer Perceptron)를 반복적으로 호출해야 한다. 이 과정은 장면 전체에 걸쳐 수백만 개 이상의 연산을 요구하므로 학습 시 많은 시간이 소요될 뿐 아니라, 고주파 성분이 충분히 샘플링 되지 못해 aliasing 현상이 발생하는 한계도 있다. 이로 인해 렌더링 시에도 수 초에서 수 분 단위의 계산이 필요하여 학습 및 렌더링 속도 측면에서 비효율적이다.

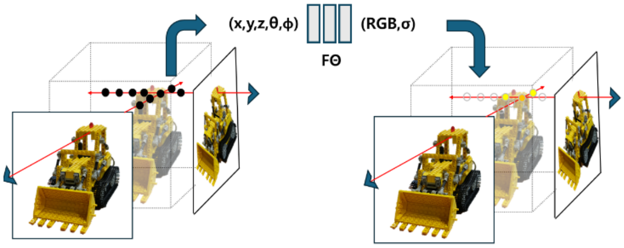

그림 1. NeRF의 기본 구조. 장면의 3차원 좌표와 해당 지점의 시점 방향을 입력으로 받아 MLP가 각 샘플 지점의 색상 값과 밀도를 예측한다. 카메라에서 투영된 광선은 이러한 예측 값을 따라 연속적으로 적분되며 누적 결과가 픽셀 단위로 합성되어 최종 영상이 생성된다.

Fig. 1. The basic structure of NeRF. 3D scene coordinates and the corresponding viewing directions are provided as inputs to a multilayer perceptron, which predicts color and density for each sampled point. Rays cast from the camera accumulate these predictions through continuous volumetric integration, and the aggregated results from the pixel values of the final rendered image.

1.2 Mip-NeRF

NeRF는 새로운 시점의 이미지를 합성하는 데 성공했지만, 고주파 정보를 포함한 복잡한 장면을 학습하거나 낮은 해상도로 렌더링할 때 고주파 성분이 충분히 샘플링되지 못해 저주파 패턴으로 왜곡되는 aliasing과 세부 정보 손실로 인한 blur 현상이 동시에 나타나는 한계가 있었다.

2021년에 발표된 Mip-NeRF는 NeRF의 구조를 기반으로 하면서도 다중 해상도 데이터에서 발생하는 aliasing 문제를 효과적으로 처리하기 위해 제안된 모델이다. Mip-NeRF는 기존의 포인트 단위 샘플링을 대체하기 위해 cone tracing 기반 구간 적분 방식을 도입하였다.

그림 2는 cone tracing의 시각화이다. 기존 NeRF는 단일 점만 샘플링해 넓은 공간 정보를 근사하기 때문에 고주파 성분이 저주파 패턴으로 치환되는 aliasing 문제가 발생한다. 반면 Mip-NeRF는 픽셀을 한 점이 아니라 반지름 $\dot{r}$을 갖는 원형 영역으로 간주한다. 따라서 ray는 깊이 구간 $[t_0, t_1]$에 걸쳐 진행하면서 단순 선분이 아니라 밑면 반지름 $\dot{r}$을 가진 원뿔을 형성한다. 그리고 원뿔 구간 전체를 적분하고, 해당 영역을 영상 평면에 타원으로 투영하여 평균적인 색상과 밀도 값을 집계한다. 이를 통해 주파수 손실을 줄이고 멀리 있는 물체나 다운샘플링된 장면에서 뚜렷하게

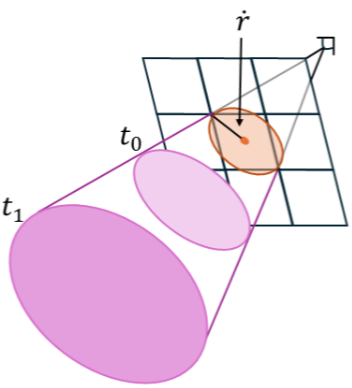

그림 2. Mip-NeRF의 cone tracing 과정. 카메라에서 발사된 ray는 깊이 구간 [t_0, t_1]에서 원뿔 모양의 부피로 확장된다. 이 부피는 영상 평면 상에서 2차원 ellipse로 투영되며, 해당 영역의 색상과 밀도는 평균적으로 집계된다.
Fig. 2. Cone tracing process of Mip-NeRF. A ray emiitted from the camera expands into a conical frustum over the depth interval [t_0, t_1]. This frustum is projected as a 2D ellipse on the image plane, where the color and density within the region are aggregated in an averaged manner.

타나는 aliasing 현상을 효과적으로 완화한다. 또한 새로운 좌표 표현 방식인 Mip representation을 도입하였다. 이 표현은 입력 3D 좌표를 단일 해상도로 인코딩하지 않고 다중 해상도의 공간에 매핑하여 각 해상도별로 정보를 학습하게 설계하였다. 이로써 작은 물체의 세밀한 구조에서 큰 물체의 전역적 패턴까지 다양한 스케일의 특징을 동시에 학습할 수 있다. 결과적으로 multi-scale 학습과 anti-aliasing을 동시에 달성하였다.

그러나 Mip-NeRF는 NeRF의 구조를 유지하기 때문에 여전히 학습 및 렌더링 속도가 느리고 복잡한 신경망 연산과 다중 해상도 처리가 추가적인 계산 부담을 야기한다는 한계가 있다. 따라서 aliasing 문제는 완화되었지만 실시간 응용에는 여전히 제약이 있다.

1.3 Instant-NGP

기존 NeRF는 뛰어난 장면 재구성 성능에도 불구하고, 방대한 연산량과 긴 학습 시간으로 인해 대규모 데이터셋이나 실제 응용에 제약이 존재한다.

이러한 한계를 극복하기 위해 2022년에 Instant-NGP가 제안되었다. Instant-NGP는 NeRF의 구조적 한계를 직접적으로 변경하기보다는 효율적인 데이터 표현 방식과 학습 최적화 기법을 통해 학습 속도와 렌더링 속도를 획기적으로 향상시켰다. 기존 NeRF는 입력 좌표의 고주파 성분을 학습하기 위해 sine과 cosine 함수를 이용한 positional encoding을 사용했는데 고해상도 세부 표현을 위해 주파수 레벨을 높일수록 입력 차원이 증가하여 MLP 파라미터 수와 연산량이 선형적으로 증가하는 문제가 있었다. 반면 Instant-NGP의 핵심인 multi-resolution hash encoding은 입력 좌표를 여러 해상도의 격자에 매핑하고, 이를 해시 테이블 구조에 저장하여 동일한 격자 공간을 효율적으로 공유하도록 설계되었다.

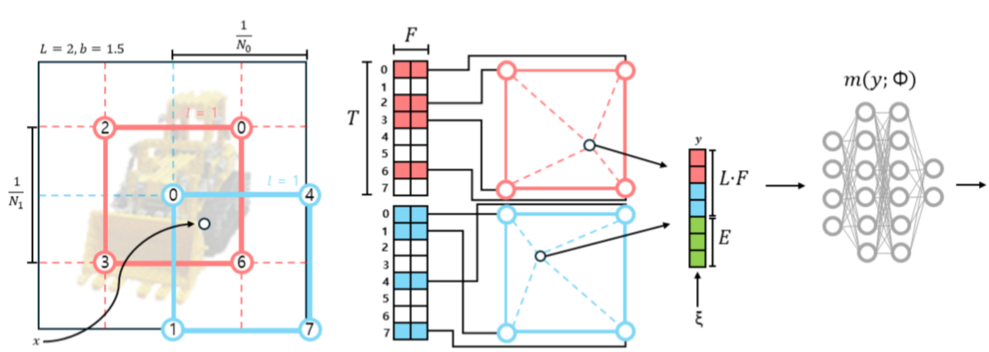

그림 3. Instant-NGP의 multi-resolution hash encoding 구조. 입력 좌표는 다중 해상도의 격자 공간에 매핑되며 각 해상도마다 인접 격자의 feature가 해시 테이블을 통해 불러와진다. 이러한 feature들은 보간 과정을 거쳐 연결된 후, 최종적으로 MLP에 전달되어 색상과 밀도를 예측한다.
Fig. 3. Multi-resolution hash encoding in Instant-NGP. The input coordinate is mapped onto grids at multiple resolutions, and the features of neighboring grid vertices are retrieved via hash tables. These features are interpolated and concatenated across levels, and the aggregated representation is fed into an MLP to predict color and density.

그림 3과 같이 입력 좌표는 다중 해상도의 격자 공간에 사상되고, 해당 격자의 인접 정점들이 해시 테이블에 의해 관리된다. 각 정점은 해시 함수로 해시 테이블의 특정 위치에 매핑되며, 이때 해시 테이블은 크기 T와 각 정점당 feature 차원 F를 가진다. 매핑된 위치에서 정점의 feature 벡터가 조회되고, 이렇게 얻은 feature들은 좌표의 상대적 위치를 고려하여 선형 보간되어 하나의 연속적 표현으로 합쳐진다. 이후 레벨별 결과가 모두 연결되어 $L \cdot F$ 차원의 벡터가 형성되며, 필요한 경우 추가 입력 E와 결합된다. 마지막으로 이 벡터는 신경망 $m(y;\Phi)$의 입력으로 들어가 색상과 밀도를 출력한다. 이 방식은 필요한 파라미터 수를 크게 줄이면서 장면 내 복잡한 세부 정보를 효율적으로 학습할 수 있다. 또한 CUDA 기반 병렬 최적화는 GPU의 쓰레드 단위를 세밀하게 활용하여 각 ray 샘플에 대한 연산을 병렬 처리할 수 있도록 하였고, mixed-precision training은 부동소수점 연산을 FP32 대신 FP16으로 대체하여 연산 속도를 높이고 메모리 대역폭을 절감하였다. 그 결과 기존 NeRF에서 수 시간 이상 소요되던 학습을 수 분 단위로 단축시킬 수 있었으며, 복잡한 장면에서도 실시간 렌더링이 가능해졌다.

그러나 Instant-NGP는 해시 테이블 구조가 좌표를 제한된 크기의 버킷에 매핑하기 때문에, 복잡한 기하 구조에서는 서로 다른 위치 정보가 동일한 해시 버킷을 공유하는 충돌이 빈번하게 발생할 수 있다. 결과적으로 매우 복잡하거나 비정형적인 장면에서는 최적화가 어렵다는 한계가 존재한다. 또한 기본적인 표현력은 여전히 NeRF 구조에 기반하고 있기 때문에 aliasing이나 일반화 문제를 근본적으로 해결하지는 못했다.

2. 3DGS based model

2023년에 제안된 3DGS는 장면을 수많은 3차원 Gaussian 집합으로 직접 표현하는 방식이다. 이는 포인트 클라우드 기반 표현을 확장한 것으로, 각 Gaussian이 겹쳐 장면을 형성한다. 이후 렌더링 결과와 GT (Ground Truth)를 비교해 Gaussian의 파라미터를 최적화한다.

2.1 Vanilla 3DGS

그림 4는 3DGS의 전체 파이프라인을 나타낸다. 3DGS는 장면을 MLP를 기반으로 암묵적으로 표현하는 대신, 각 장면 요소를 수천~수만 개의 3차원 Gaussian으로 구성하며, 각 Gaussian은 위치 좌표, 방향 벡터, 크기, 색상, 불투명도 파라미터를 지닌다. 이러한 Gaussian은 초기 입력으로 사용되는 포인트 클라우드로부터 생성된다. 해당 포인트 클라우드는 SfM (Structure-from-Motion) 기법을 통해 복원되며, SfM은 다수의 영상 프레임 간의 대응 관계를 바탕으로 카메라 포즈 및 공간 상의 3D 점들을 추정하는 구조 복원 기술이다[11]. SfM으로부터 얻은 포인트 클라우드는 Initialization 단계를 통해 각 포인트가 하나의 Gaussian으로 변환된다. Gaussian 파라미터는 포인트 클라우드의 기하학적 정보와 색상 정보를 기반으로 초기화되며, 이후 학습을 통해 지속적으로 업데이트된다. 변환된 Gaussian은 Projection 단계를 통해 카메라 시점에 맞춰 2차원 이미지 평면으로 투영된다. 이때 카메라의 내·외부 파라미터를 고려하여 각 Gaussian이 이미지 평면 상의 어떤 위치와 크기, 형태로 표현될지를 계산하고, 그 결과는 다음 단계인

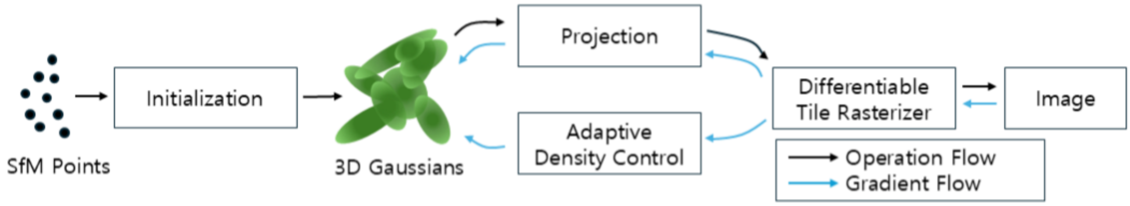

그림 4. 3DGS의 파이프라인. SfM으로 추출된 포인트를 초기화하여 3차원 Gaussian으로 장면을 표현한다. 이후 Projection과 Adaptive Density Control을 거쳐 Differentiable Tile Rasterizer에서 렌더링이 수행되며 최종적으로 이미지를 생성한다. 검은색 화살표는 연산 흐름을, 파란색 화살표는 최적화 흐름을 나타낸다.

Fig. 4. The pipeline of 3DGS. Points reconstructed by SfM are initialized and represented as 3D Gaussians. These are processed through projection and adaptive density control, followed by Differentiable Tile Rasterization to generate the final image. Black arrows denote the operation path, while blue arrows indicate the gradient propagation path.

Differentiable Tile Rasterizer로 전달된다. Rasterizer는 각 Gaussian이 이미지 내에서 차지하는 영역에 대해 연산을 수행하고 최종 이미지를 합성한다. 이 연산은 미분 가능한 구조로 설계되어 손실 계산 및 역전파를 가능하게 만드는 핵심 요소이다. 합성된 이미지는 원본 이미지와 비교되어 손실이 계산되며, 이 손실은 역전파되어 각 Gaussian의 위치, 색상, 크기, 불투명도 파라미터에 gradient를 전달함으로써 전체 장면 표현이 최적화된다.

그림 5와 같이 Adaptive Density Control을 통해 Gaussian 배치를 동적으로 재구성한다. 우선 일정 간격으로 각 Gaussian의 positional gradient를 산출하고 임계값을 초과한 Gaussian을 선정한다. 이때 특정 영역이 충분히 재구성되지 않아 디테일이 부족한 경우를 under-reconstruction이라 하며, 이러한 영역에는 Gaussian을 복제하여 더 많은 점을 배치한다. 반대로 특정 영역에 과도하게 밀집되어 불필요하게 Gaussian이 낭비되는 경우를 over-reconstruction이라 하며, 이러한 영역에서는 기존 Gaussian을 분할하여 더 작은 단위로 나눈다. 이와 같은 방식은 영역별 밀도를 균형 있게 유지하면서 최적화를 지속할 수 있도록 한다. 또한 신경망 기반 접근보다 훨씬 빠른 렌더링 속도를 제공하고 고해상도 장면에 대한 정밀한 묘사가 가능하다.

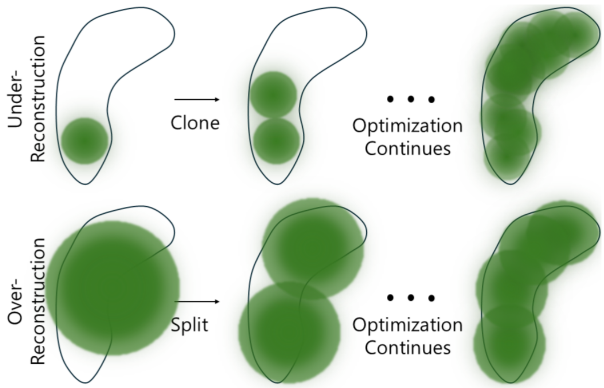

그림 5. 3DGS의 Adaptive Density Control. Under-Reconstruction이 발생한 영역에서는 Gaussian을 복제(clone)하여 부족한 디테일을 보완하고, Over-Reconstruction이 발생한 영역에서는 Gaussian을 분할(split)하여 과도하게 표현된 영역을 세분화한다.

Fig. 5. Adaptive density control process of 3DGS. In Under-Reconstruction regions, Gaussians are cloned to compensate for missing details, while in Over-Reconstruction regions, Gaussians are split to refine overly large representations.

하지만, 3DGS는 Gaussian 수가 많아질수록 GPU 메모리 사용량이 급격히 증가한다. 그리고 Gaussian 초기화가 SfM 기반 포인트 클라우드에 전적으로 의존하기 때문에 부정확한 초기 점들은 학습 전체에 누적되어 오류를 야기할 수 있다. 그래서 초기 포인트 클라우드 품질에 따라 최종 재구성 결과가 크게 좌우된다. 그리고 멀리 있는 객체를 렌더링하거나 낮은 해상도 환경에서는 고주파 성분을 충분히 반영하지 못해 Gaussian 투영 시 장면이 흐려지거나 세부 디테일 손실이 발생하는 한계가 있다.

2.2 Mip-Splatting

3DGS는 장면을 연속적인 복셀이나 볼륨으로 표현하지 않고, Gaussian으로 구성함으로써, 학습과 렌더링 속도에서 큰 장점을 보였다. 그러나 멀리 있는 객체나 해상도가 낮은 환경에서 렌더링할 경우, blur와 디테일 손실이 발생할 수 있다.

이를 개선하기 위해 Mip-Splatting이 제안되었다. Mip-Splatting은 3DGS의 효율적인 구조를 유지하면서도 NeRF 계열의 Mip-NeRF에서 도입된 cone tracing 기반 적분 방식을 Gaussian Splatting에 적용하였다. 각 Gaussian 분포를 단순히 픽셀 단위로 투영하지 않고 cone tracing 기반 적분 방식으로 처리함으로써, 단일 픽셀 내 여러 고주파 성분이 단일 점 샘플링에서 aliasing으로 왜곡되는 문제를 방지하고, 해당 구간의 평균 분포를 반영하여 실제 장면에 더 근접한 결과를 제공한다. 또한 Mip representation을 도입하여 Gaussian을 단일 해상도로 고정하지 않고, 객체의 크기, 거리, 시야각에 따라 표현 스케일을 동적으로 조정할 수 있도록 설계되었다. 각 Gaussian은 다중 스케일로 정의되어 가까운 영역에서는 작은 커널로 세밀한 구조를 유지하고, 먼 영역이나 다운샘플링된 상황에서는 더 넓은 커널로 통합된다. 이렇게 다층적 표현을 적용하면 한 장면 안에서도 서로 다른 스케일의 정보를 동시에 학습할 수 있다. 이러한 접근은 3DGS 대비 고주파 영역의 세부 묘사가 개선되었으며 저해상도 렌더링 상황에서도 안정적인 시각적 품질을 유지할 수 있다.

그러나 Mip-Splatting 역시 Gaussian 기반 구조의 특성상 장면의 복잡도가 커질수록 Gaussian 수가 급격히 증가하여 GPU 메모리 자원을 크게 소모하게 된다. 또한 초기

Gaussian은 여전히 포인트 클라우드 품질에 의존한다. 그 결과, 초기 재구성이 부정확하거나 잡음이 많을 경우 잘못 배치된 Gaussian이 학습 과정 전반에 걸쳐 누적되어 왜곡된 결과를 초래할 수 있다.

2.3 2DGS

기존의 3DGS는 빠른 학습과 고속 렌더링을 동시에 달성하였다. 그러나 Gaussian이 표면보다 앞뒤로 퍼져서 배치될 경우, 일부 영역은 겹치거나 비어 있게 된다. 그 결과, 표면의 세부 구조가 왜곡되어서 기하학적 정확도 측면에서 한계를 가진다. 특히 복잡한 형상일수록 이러한 불일치가 누적되어 정밀한 재구성이 제한된다.

이러한 한계를 해결하기 위해 2024년에 2DGS가 제안되었다. 2DGS는 장면의 기하학적 구조를 더 정확히 반영하기 위해 각 Gaussian을 3차원 볼륨 분포가 아닌, 표면에 정렬된, 타원 디스크 형태의 2D Gaussian primitive로 정의한다.

그림 6과 같이 2DGS는 기존 3DGS에서 사용하던 3차원 분포를 직접 투영하는 방식 대신, 이미지 평면 위에 2차원 Gaussian을 정의하여 렌더링 과정을 단순화 한다. 3D Gaussian은 카메라 광선과 교차하는 평면에서 타원 형태로 투영되어야 하므로 복잡한 3차원 계산이 필요하다. 반면 2DGS는 처음부터 이미지 평면 위에 Gaussian 분포를 배치함으로써 교차 연산을 거치지 않고도 픽셀 단위에서 직접 밀도와 색상을 계산할 수 있다. 그리고 2차원으로 제한하면 각 Gaussian이 실제 물체 표면의 법선 방향과 일관되도록 정렬되므로 불필요하게 부피 공간을 차지하지 않고 실제 형상에 밀착된 표현이 가능하다. 그 결과, 표면에서의 기하학적 왜곡이 줄어들고 세부 구조의 충실도가 크게 개선되었다.

그러나 Gaussian을 2차원으로 제안함으로써 표면 정렬에 유리하지만, 부피적 표현 능력이 사라져 안개나 연기같은 비표면 기반 현상이나 장면 내부 구조를 포함한 복잡한 3차원 기하를 유연하게 표현하는 데에는 제약이 존재한다. 그리고 여전히 초기 포인트 클라우드 품질에 영향을 받는다는 점에서 한계가 남아 있다.

III. 연구 방법

본 연구에서는 NeRF 계열 모델을 각각 100K iteration까지, 3DGS 계열 모델은 각각 7K iteration 및 30K iteration까지 학습을 진행하였다. 이는 각 모델에서 기존 연구가 제시한 안정적인 성능 지점을 기준으로 설정하였다. 데이터셋은 T&T (Tanks & Temples) 데이터셋의 Francis, Museum과 LLFF (Local Light Field Fusion) 데이터셋인 Bonsai를 사용하여 실험을 진행하였다. 데이터셋은 전체 이미지에서 8번째 이미지마다 테스트셋으로 분류하고, 나머지를 학습용 트레이닝셋으로 사용하였다. 성능 평가는 PSNR, SSIM, LPIPS와 같은 시각 품질 지표와 Training Time, Training Peak Memory를 포함한 연산 효율성 지표를 사용하여 평가하였다. 시각 품질 지표는 테스트셋 전체

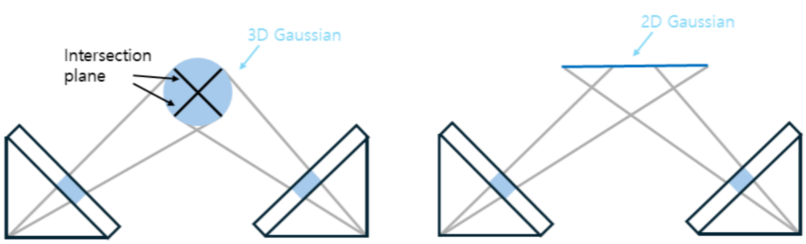

그림 6. 3D Gaussian과 2D Gaussian의 투영 방식 비교. 3D Gaussian은 카메라에서 발사된 ray와 교차 평면을 통해 이미지 평면에 투영되며, 그 결과 타원 형태의 footprint가 형성된다. 반면 2D Gaussian은 이미지 평면 상에 직접 정의되어, 복잡한 3차원 교차 과정을 거치지 않고 바로 2차원 분포로 표현된다.

Fig. 6. Comparison of projection between 3D and 2D Gaussian. A 3D Gaussian intersects with the camera rays on an image-aligned plane, producing an elliptical footprint on the image plane. A 2D Gaussian is directly defined on the image plane, avoiding explicit 3D intersections and simplifying the representation and rendering process.

에 대해 평균값을 산출하여 결과의 신뢰성을 확보하였다. 또한 정량적 평가뿐만 아니라, 각 모델이 생성한 렌더링 이미지를 제시하고 화질과 질감 표현을 중심으로 정성적 분석을 수행하였다. 이를 통해 수치화된 지표뿐만 아니라 시각적으로 인지되는 품질 차이까지 종합적으로 확인할 수 있다.

1. 실험 환경

표 1은 본 논문에서 모델을 비교 평가하기 위해 사용된 실험 환경을 정리한 것이다. 각 모델은 Intel Core i7-13700 CPU, NVIDIA RTX 4080 GPU, 64GB RAM으로 구성된 동일한 하드웨어 환경에서 학습 및 테스트를 수행하였다. 소프트웨어 환경은 각 모델이 구현된 프레임워크 버전 및 의존성을 고려해 다르게 설정하였다. 구체적으로, NeRF와 3DGS는 Pytorch 2.1.2 기반으로 구현되어 있어 동일한 버전으로 실험을 진행하였다. Mip-Splatting은 공식 구현에서 PyTorch 2.5.1 환경을 요구하므로 해당 버전을 적용하였으며, 2DGS는 Pytorch 2.0.0 기반에서 안정적인 학습이 가능하도록 최적화되어 있어 그에 맞추어 설정하였다. 또한 CUDA 역시 각 프레임워크와 드라이버 호환성을 반영하여 11.8, 12.1, 12.3 버전을 선택하였다. 이와 같이 소프트웨어 버전이 상이한 이유는 각 모델의 오픈소스 구현체에서 제공하는 권장 환경을 따름으로써 재현 가능성을 확보하기 위함이다. 즉, 임의로 통일된 버전을 강제하기보다는 각 모델의 성능이 본래 의도된 환경에서 발휘되도록 설정하여 비교의 신뢰성을 높였다.

2. 평가 지표

3D 장면 재구성 및 새로운 시점 영상 합성 기술의 성능을 비교·평가하기 위해서는 재구성된 이미지가 원본 이미지와 얼마나 유사한지를 판단할 수 있는 정량적이고 객관적인 지표가 요구된다. 컴퓨터 비전 및 컴퓨터 그래픽스 분야에서는 PSNR, SSIM, LPIPS와 같은 지표들이 대표적으로 사용된다. 이러한 평가 지표들은 단순한 화소 간 차이뿐만 아니라, 구조적 유사성이나 시각적 품질 차이 등을 반영함으로써 실제 사용자 관점에서의 품질 차이를 보다 정확히 평가할 수 있게 해준다.

PSNR은 재구성된 이미지와 원본 이미지 간의 픽셀 단위 차이를 기반으로 정의되며, MSE (Mean Squared Error)에 기초한다. PSNR은 최대 픽셀값 MAX_I와 MSE를 이용해 식 (1)과 같이 정의된다.

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (1)$$

여기서 MSE는 두 영상의 픽셀 차이를 제곱 평균한 값이다. PSNR값이 클수록 원본과의 유사성이 높고 복원 품질이 우수함을 의미한다.

SSIM은 영상 내 밝기, 대비, 구조 정보를 종합적으로 고려하여 두 영상 간의 구조적 유사도를 측정하는 지표이다. SSIM은 원본 이미지인 x와 재구성된 이미지인 y 사이에서 식 (2)와 같이 정의된다.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

표 1. 실험 모델에 사용된 하드웨어 및 소프트웨어 환경. 본 표는 모든 모델의 공통 하드웨어 환경과, 모델별 소프트웨어 환경을 함께 정리하였다.
Table 1. Hardware and software environments used for the experiments. The table summarizes the common hardware configuration and the model-specific software settings.

		NeRF	Mip-NeRF	Instant-NGP	3DGS	Mip-splatting	2DGS
HardWare	CPU	Intel Core i7-13700					
	GPU	NVIDIA RTX 4080					
	RAM	64GB					
SoftWare	Ubuntu	22.04.3 LTS					
	PyTorch	2.1.2				2.5.1	2.0.0
	CUDA	11.8			12.3	12.1	11.8

여기서 μ_x, μ_y는 두 영상의 평균 밝기, σ_x, σ_y는 표준편차, σ_{xy}는 공분산을 의미한다. C_1, C_2는 분모의 안정성을 보장하기 위한 작은 상수이다. SSIM의 값이 1에 가까울수록 원본과 구조적 유사성이 높음을 의미하며 값이 작을수록 구조 왜곡이나 노이즈가 많음을 나타낸다.

LPIPS는 두 영상의 시각적 유사도를 평가하기 위한 지표로 사람이 인지하는 지각적 차이를 정량화한다. LPIPS는 고차원 이미지 표현 학습 기반의 딥러닝 모델을 활용하여 실제 인간의 지각에 가까운 시각적 유사도를 평가한다. LPIPS는 입력 영상 x,y를 신경망에 통과시킨 후, 각 계층 l의 특징 맵 $\hat{x}^l_{h,w}$, $\hat{y}^l_{h,w}$을 정규화하여 계산된 거리의 가중합으로 식 (3)과 같이 정의된다.

$$LPIPS(x,y) = \sum_l w_l \cdot (1/(H_l W_l)) \sum_{h,w} \|\hat{x}^l_{h,w} - \hat{y}^l_{h,w}\|_2^2 \quad (3)$$

LPIPS값은 작을수록 두 영상이 시각적으로 유사함을 의미하며, 값이 클수록 시각적 차이가 큼을 나타낸다.

3. Datasets

본 연구에서는 Novel View Synthesis 분야에서 널리 사용되는 대표적인 공개 벤치마크 데이터셋인 T&T 및 LLFF 데이터셋을 활용하였다[12][13]. 구체적으로는 Francis, Museum, Bonsai 세 가지 장면을 실험 대상으로 선정하였다. T&T 데이터셋인 Fraincis는 실외 공간의 다양한 깊이감을 포함한다. 실외 공간의 재구성 성능을 검증하기 위해 자주 사용된다. 그리고 T&T 데이터셋인 Museum은 실내 공간의 조명 변화와 세밀한 질감을 포함한다. 복잡한 시각적 요소에 대한 모델의 일반화 능력을 평가하는 데 자주 사용된다. 다음으로 LLFF 데이터셋인 Bonsai는 실내에서 오브젝트를 중심으로 촬영된 장면이다. 오브젝트의 세밀한 구조를 포함하고 있어, 고주파 디테일 복원 성능을 평가하는 데 사용된다.

Ⅳ. 실험 결과 및 분석

1. 정량적 평가

표 2는 Francis, Museum, Bonsai 데이터셋에서의 정량적 성능 평가 결과를 보여주며 각 모델의 구조적 특성과 성능 간의 연관성을 분석한다.

먼저, NeRF는 세 장면 모두에서 가장 낮은 성능을 기록하였다. PSNR은 약 12dB, SSIM 0.27~0.41, LPIPS는 0.8이상으로 높게 나타났다. 이는 NeRF가 모든 위치·방향 쿼리를 단순한 MLP로 직접 학습하기 때문에 고해상도 및 복잡한 기하 구조를 표현하는 데 비효율적이기 때문이다. 또한 NeRF는 aliasing 현상에 취약하여 제한된 학습으로는 고품

표 2. Francis, Museum, Bonsai 데이터셋에서 NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, 2DGS 모델의 PSNR, SSIM, LPIPS 비교. 각 모델의 화질 및 시각적 유사도를 객관적으로 비교할 수 있다. 표 내 색상은 성능 수준을 강조하기 위한 것으로, 빨강은 최상위 결과, 주황은 차상위 결과, 노랑은 차차상위 결과를 나타낸다.
Table 2. Comparison of PSNR, SSIM, and LPIPS for NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, and 2DGS on the Francis, Museum, and Bonsai datasets. Colors highlight the performance levels: red denotes the top results, orange the second-best, yellow the third-best results.

Dataset	Francis			Museum			Bonsai		
Method|Metrics	PSNR [dB]↑	SSIM↑	LPIPS↓	PSNR [dB]↑	SSIM↑	LPIPS↓	PSNR [dB]↑	SSIM↑	LPIPS↓
NeRF	12.290	0.414	0.851	12.202	0.273	1.042	12.467	0.385	0.837
Mip-NeRF	12.674	0.412	0.795	12.131	0.262	0.948	12.725	0.387	0.847
Instant-NGP	23.571	0.803	0.149	25.824	0.931	0.040	27.530	0.919	0.095
3DGS 7K	28.919	0.898	0.187	33.085	0.955	0.056	29.715	0.924	0.214
3DGS 30K	32.434	0.921	0.151	34.817	0.967	0.044	31.960	0.942	0.182
Mip-Splat 7K	31.058	0.924	0.145	34.718	0.966	0.048	30.621	0.937	0.192
Mip-Splat 30K	34.175	0.941	0.117	37.130	0.976	0.037	34.010	0.959	0.157
2DGS 7K	28.165	0.892	0.192	32.513	0.948	0.061	29.391	0.921	0.220
2DGS 30K	30.482	0.908	0.174	33.442	0.954	0.054	31.358	0.935	0.205

질 영상을 복원하기 어렵다.

Mip-NeRF는 aliasing을 고려한 mip representation을 도입하여 다중 해상도의 시그널을 안정적으로 학습한다. 그러나 여전히 MLP를 기반으로 한 느린 최적화와 제한된 표현력 때문에 시각 품질의 개선 폭은 제한적이었다. 또한 Mip-NeRF는 샘플링 단계에서의 왜곡은 줄였지만, 복잡한 장면의 세밀한 재현에는 근본적 한계가 존재한다.

Instant-NGP는 hash encoding을 이용한 효율적인 위치 인코딩 기법을 통해 장면 내 고주파 정보를 효율적으로 학습할 수 있다. 그 결과 PSNR은 Francis 23.575dB~27.503dB로 크게 상승하였고, SSIM도 0.803~0.931로 높은 성능을 달성하였다. 특히 LPIPS가 0.149 이하로 낮게 나타난 것은 hash encoding이 지역적 구조와 세부 텍스처를 효과적으로 포착하여, 사람이 인지하는 화질 품질을 잘 보존했음을 의미한다. 그러나 최적화 단계가 다른 모델들에 비해 비교적 얕기 때문에 최종 정밀도는 낮은 편이다.

3DGS는 연속적 장면 표현을 MLP가 아닌 Gaussian 분포 집합으로 치환하여 직접 최적화한다. 이 방식은 학습 속도가 빠르고, 시각적으로 연속적인 표면 표현이 가능하다. 실제로 7K Iteration만으로 Instant-NGP보다 높은 PSNR과 SSIM을 달성하였으며, 30K Iteration에서는 Francis 32.434dB, Museum 34.817dB, Bonsai 31.960dB로 높은 정밀도를 보였다. 다만 LPIPS가 Francis 0.151, Bonsai 0.182 수준으로, 이는 Gaussian이 고주파 디테일을 완벽히 보존하지 못하며 실내 공간 재구성 성능이 부족한 결과로 해석할 수 있다.

Mip-Splatting은 세 장면에서 모두 가장 높은 성능을 기록하였다. Mip-Splatting은 3DGS에 다중 해상도 특성을 결합하여 aliasing 문제를 줄이고, 복잡한 기하의 고주파 성분을 보다 안정적으로 학습하도록 개선한 방식이다. 그 결과, PSNR이 Francis 34.175dB, Museum 37.130dB, Bonsai 34.010dB로 가장 높았으며, SSIM도 0.941~0.976으로 최고 수준을 기록하였다. LPIPS는 0.037~0.157를 달성하며 구조적 정확성과 지각적 유사도를 동시에 달성하였다. 이는 Mip-Splatting이 복잡한 장면에서 발생하는 multi-scale aliasing 문제를 효과적으로 해결했음을 보여준다.

2DGS는 Gaussian 표현을 3차원이 아닌 2차원 이미지 평면으로 제안하여 계산을 단순화한다. 이로 인해 깊이 축 정보를 직접적으로 모델링하지 못해 3DGS나 Mip-Splatting 대비 정밀도가 낮다. 실제로 PSNR은 30.482dB~33.442dB, SSIM은 0.908~0.954, LPIPS가 0.174~0.205로 3DGS 계열 모델 중에서는 시각 품질에서 낮은 성능을 보였다. 따라서 2DGS는 고품질 재구성에는 제약이 존재한다.

표 3은 Francis, Museum, Bonsai 데이터셋에서 각 모델의 학습 시간과 GPU 메모리 사용량을 비교한 결과이다.

NeRF와 Mip-NeRF는 모든 장면에서 긴 학습 시간이

표 3. Francis, Museum, Bonsai 데이터셋에서 NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, 2DGS 모델의 학습 시간과 메모리 사용량 비교. 짧은 훈련 시간과 낮은 메모리 사용량은 효율적인 학습을 의미하며, 이를 통해 각 모델의 연산 효율성과 자원 소모를 객관적으로 비교할 수 있다. 표 내 색상은 효율성 수준을 강조하기 위한 것으로, 빨강은 최상위 결과, 주황은 차상위 결과, 노랑은 차차상위 결과를 나타낸다.
Table 3. Comparison of training time and training peak memory consumption of NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, and 2DGS on the Francis, Museum, Bonsai datasets. Shorter training time and lower memory consumption indicate higher efficiency, providing an objective comparison of computational cost across models. Colors highlight the efficiency levels: red denotes the top results, orange the second-best, and yellow the third-best results.

Dataset	Francis		Museum		Bonsai	
Method|Metrics	Training Time	Memory	Training Time	Memory	Training Time	Memory
NeRF	106m41s	5.50GB	106m29s	5.11GB	110m2s	5.17GB
Mip-NeRF	108m5s	5.55GB	107m39s	5.16GB	108m36s	5.16GB
Instant-NGP	22m5s	4.34GB	22m22s	3.97GB	27m44s	3.86GB
3DGS 7K	1m13s	2.44GB	1m31s	3.87GB	3m33s	10.18GB
3DGS 30K	5m12s	2.53GB	7m40s	4.12GB	15m3s	10.18GB
Mip-Splat 7K	1m48s	1.78GB	2m30s	2.43GB	5m6s	8.61GB
Mip-Splat 30K	8m14s	1.94GB	13m5s	2.65GB	24m41s	8.77GB
2DGS 7K	1m51s	0.83GB	2m22s	1.47GB	7m2s	3.67GB
2DGS 30K	7m55s	1.84GB	11m2s	2.69GB	30m22s	14.27GB

소요되었으며, GPU 메모리 사용량은 5GB 수준이었다. 이는 단일 MLP 기반 구조가 모든 위치·방향 쿼리를 순차적으로 최적화하기 때문에 학습 효율이 낮음을 보여준다. Mip-NeRF는 anti-aliasing 처리를 포함하였으나, 학습 시간과 메모리 사용량에서 NeRF와 큰 차이를 보이지 않았다.

Instant-NGP는 hash encoding을 활용하여 학습 속도를 개선하면서 Francis와 Museum 데이터셋에서는 약 22분 내외로 소요되었고, Bonsai는 27분 44초가 소요되어 다른 NeRF 계열 모델 대비 크게 단축되었다. 메모리 사용량은 3.8GB~4.3GB 수준으로 다른 NeRF 계열 모델보다 효율적이었다.

3DGS는 학습 효율성에서 가장 뚜렷한 개선을 보였다. Francis에서 7K iteration은 1분 13초, 30K iteration도 5분 12초에 불과하였고, Museum에서도 7K iteration에서 1분 31초, 30K에서 7분 40초 내외로 매우 짧았다. 이는 MLP를 통한 간접 표현 대신 Gaussian 분포 집합을 직접 최적화하는 방식 덕분이다. 그러나 Bonsai와 같이 기하학적 복잡도가 높은 장면에서는 GPU 메모리 사용량이 10.18GB까지 증가하였으며, 이는 수많은 Gaussian 원소를 관리하는 과정에서 발생하는 추가 메모리 요구 때문으로 해석할 수 있다.

Mip-Splatting의 학습 시간은 Francis에서 7K iteration까지 1분 48초, 30K까지는 8분 14초로 매우 빨랐다. 학습 안정성과 효율성을 동시에 달성하였다. 이는 multi-scale splatting 구조가 메모리 사용을 보다 효율적으로 분산시키는 효과에 기인한다.

2DGS는 Francis와 Museum의 7K iteration에서는 각각 0.83GB, 1.47GB만을 사용하였으며, Bonsai에서도 3.67GB 정도 사용하였다. 이는 2DGS가 초기 학습 구간에서는 가장 높은 메모리 효율성을 보여준다. 그러나 30K iteration 학습에서는 Bonsai 장면에서 14.27GB까지 증가하여, 고주파 디테일이 높은 장면에서 메모리 사용이 급격히 늘어나는 양상을 보였다. 학습 시간은 Francis 7K iteration에서 1분 51초, Museum 7K iteration에서 2분 22초로 짧았으나, Bonsai 30K iteration에서는 30분 이상 소요되어 장면 복잡도에 따라 성능 편차가 컸다.

종합하면, NeRF와 Mip-NeRF는 학습에 수 시간이 소요되어 학습 효율성이 낮았다. Instant-NGP는 다른 NeRF 모델보다 학습 시간이 단축되었고 메모리 사용 면에서 개선이 있었다. 3DGS와 Mip-Splatting은 공통적으로 수 분 내 학습이 가능하며, 높은 품질과 효율성을 동시에 달성하였다. 특히 Mip-Splatting은 메모리 사용량을 억제하면서 안정적으로 학습을 진행할 수 있는 장점이 있었다. 2DGS는 가장 낮은 메모리 사용량을 기록했으나, 복잡한 장면에서는 학습 시간이 크게 늘어나 안정성이 떨어졌다. 따라서 학습 효율성과 품질의 균형을 고려할 때, Mip-Splatting이 가장 실용적인 대안으로 평가된다.

2. 정성적 평가

그림 7은 Francis 데이터셋에 대한 다양한 모델의 정성적 결과를 보여준다.

NeRF는 전반적으로 이미지가 흐릿하고 기하 구조가 불명확하여 세부 디테일이 거의 재현되지 않았다. 이는 단일 MLP 기반 표현이 고주파 성분 학습에 비효율적이어서, 제한된 학습 시간 내에는 정밀한 구조 복원이 어렵다는 점을 보여준다.

Mip-NeRF는 기존 NeRF에 비해 aliasing 현상을 완화하여 보다 안정적인 결과를 제공하지만, 여전히 전체적으로 선명도가 부족하고 세부 묘사력이 떨어진다.

Instant-NGP는 multi-resolution hash encoding 기반의 효율적인 표현 학습 덕분에 상대적으로 뚜렷한 경계와 배경 디테일을 뚜렷하게 복원하였으며, 다른 NeRF 계열 모델에 비해 크게 향상된 결과를 제공한다. 특히 다른 모델과 달리 사람 객체의 형태까지 비교적 안정적으로 재현하였다. 그러나 세밀한 질감 표현에서는 여전히 일부 시각적 아티팩트가 남아 있는 것을 확인할 수 있다.

3DGS의 경우, iteration 수에 따라 품질 차이가 뚜렷하였다. 7K iteration에서는 구조는 비교적 선명하나 표면 질감이 단순화하고 배경 구조가 손상되었다. 반면 30K iteration에서는 디테일이 뚜렷하게 복원되어 표면 텍스처가 GT에 근접한 수준으로 재현되었으며, 배경 구조 또한 개선되었다. 이는 3DGS는 충분한 학습 시간이 확보될 경우 높은 재구성을 달성할 수 있음을 보인다.

Mip-Splatting 역시 iteration 수에 따른 차이가 명확하다. 7K iteration에서는 안정적인 구조 복원이 가능했으나 전반

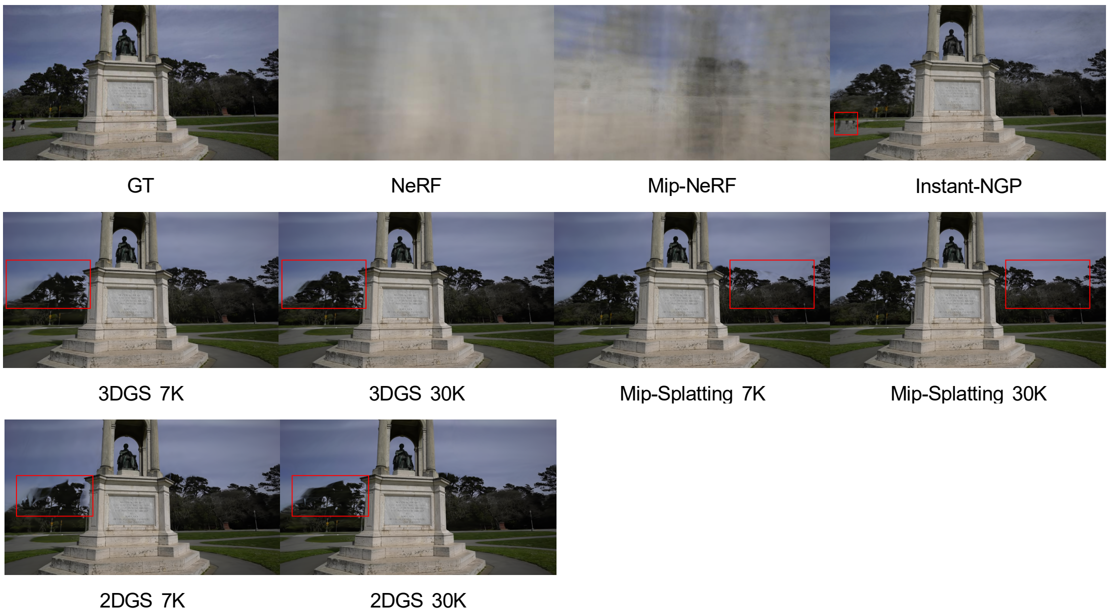

그림 7. Francis 데이터셋에서 GT와 NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, 2DGS 모델의 렌더링 결과 비교. 빨간 박스는 디테일 차이가 두드러지는 영역을 강조한 것으로, 이를 통해 각 모델 간 시각적 품질 차이를 확인할 수 있다.
Fig. 7. Visual comparison of rendered results from GT, NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, and 2DGS on the Francis dataset. Red boxes highlight regions with noticeable detail differences, enabling a qualitative comparison of visual quality across models.

적으로 디테일이 소실되어 과도하게 매끄럽게 표현되었다. 30K iteration에서는 디테일이 개선되고 배경 구조도 다른 모델보다 우수하게 재현되었다. 하지만 여전히 약간의 평활화가 관찰되었다. 이는 Mip-Splatting이 다중 해상도에서 일관성을 확보하는 데 강점이 있지만, 세밀한 구조 복원력에서 다소 한계가 있음을 시사한다.

마지막으로 2DGS는 7K iteration에서 매우 빠른 학습 속도 대비 합리적인 품질을 제공하였으나, 작은 구조와 세밀한 질감이 과도하게 평활화되어 손실되는 현상이 관찰되었다. 30K iteration에서는 전반적인 선명도가 향상되고 표면 묘사도 개선되었지만, 여전히 3DGS나 Mip-Splatting 대비 세부 기하 구조 복원 능력은 부족하였다.

그림 8은 Museum 데이터셋에 대한 다양한 모델의 정성적 비교 결과를 나타낸다.

NeRF는 영상이 전반적으로 심하게 blur 처리되어 있어, 기하 구조와 텍스처가 거의 인식되지 않는다. 이는 복잡한 실내 장면에서 NeRF가 충분한 표현력을 확보하기 어렵다는 점을 보여준다.

Mip-NeRF는 aliasing 문제를 완화하여 NeRF보다는 더 안정적인 결과를 보이지만, 여전히 세부 묘사력이 부족하고 장면 전체가 흐릿하다.

Instant-NGP는 multi-hash encoding을 통해 조명 영역을 포함한 고주파 성분을 효과적으로 학습하여, 조명 디테일이 선명하게 나타나고 텍스처도 뚜렷하게 복원되었다. 그러나 표면 질감 일부에서는 여전히 미세한 아티팩트가 관찰된다.

3DGS의 경우 7K iteration에서는 구조적 선명도가 우수했으며, 조명 형태 역시 일정 부분 복원되었으나 세밀한 빛의 질감은 단순화되어 있었다. 반면 30K iteration에서는 장식 패턴과 표면 디테일이 뚜렷하게 표현되어, GT에 근접하게 재현되었고, 빛의 질감까지 안정적으로 복원되었다.

Mip-Splatting은 7K iteration에서 구조를 잘 유지되지만 텍스처가 평활화되어 디테일이 손실되는 양상이 보인다. 30K iteration에서는 조명 테두리와 배경 패턴까지 세밀하게 표현되며, 다중 해상도 특성 덕분에 스케일 변화에도 일관된 품질을 보였다.

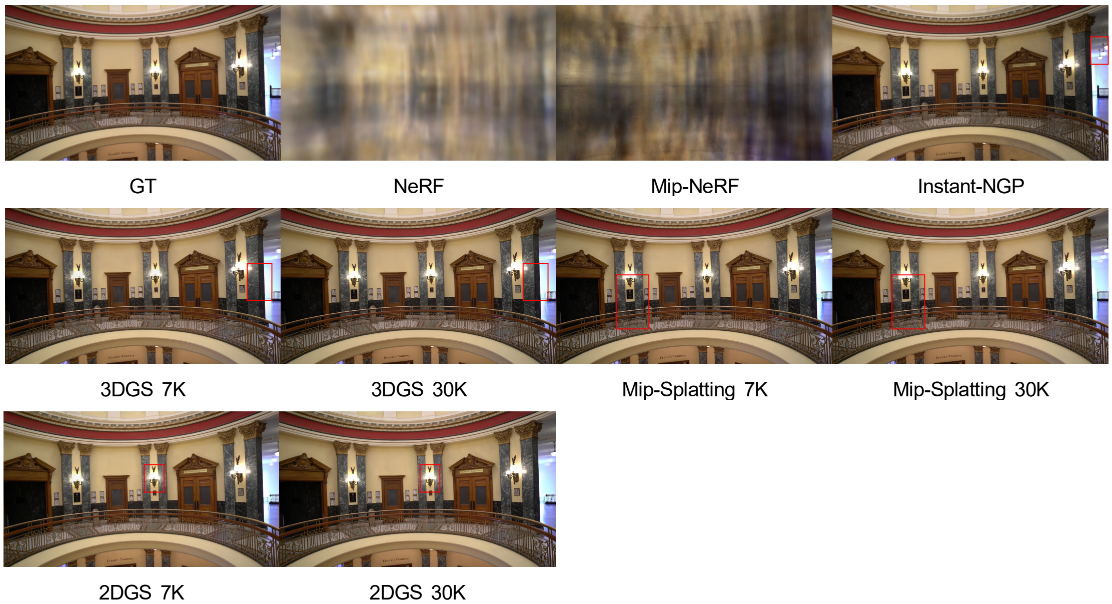

그림 8. Museum 데이터셋에서 GT와 NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, 2DGS 모델의 렌더링 결과 비교. 빨간 박스는 디테일 차이가 두드러지는 영역을 강조한 것으로, 이를 통해 각 모델 간 시각적 품질 차이를 확인할 수 있다.
Fig. 8. Visual comparison of rendered results from GT, NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, and 2DGS on the Museum dataset. Red boxes highlight regions with noticeable detail differences, enabling a qualitative comparison of visual quality across models.

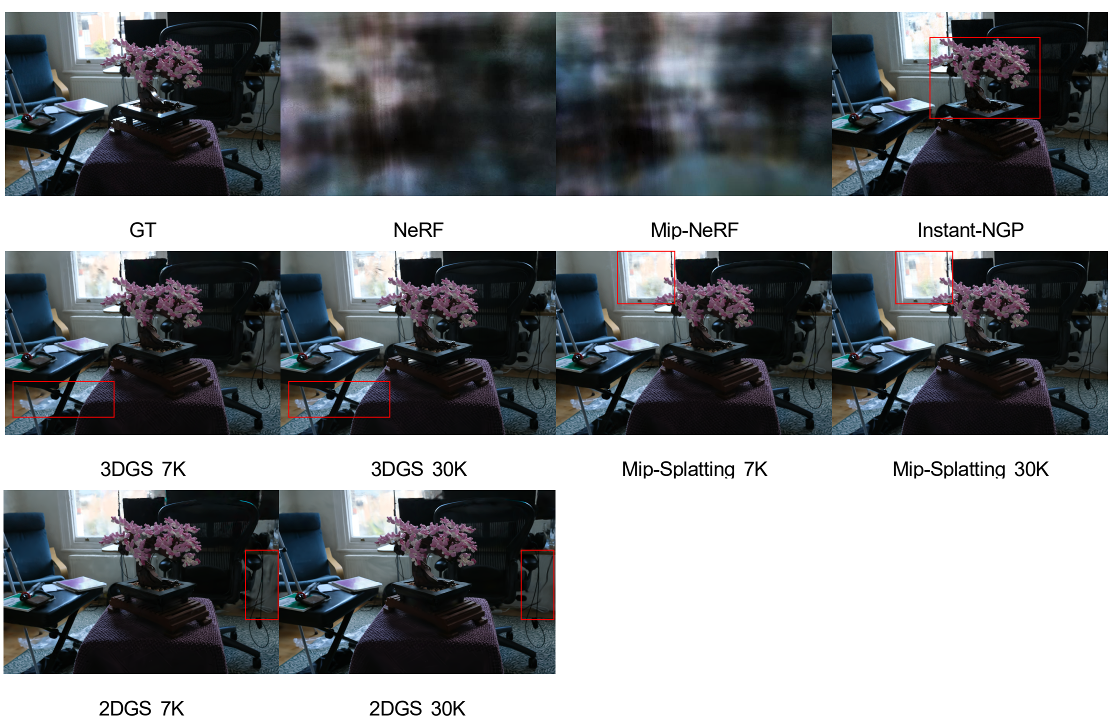

그림 9. Bonsai 데이터셋에서 GT와 NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, 2DGS 모델의 렌더링 결과 비교. 빨간 박스는 디테일 차이가 두드러지는 영역을 강조한 것으로, 이를 통해 각 모델 간 시각적 품질 차이를 확인할 수 있다.
Fig. 9. Visual comparison of rendered results from GT, NeRF, Mip-NeRF, Instant-NGP, 3DGS, Mip-Splatting, and 2DGS on the Bonsai dataset. Red boxes highlight regions with noticeable detail differences, enabling a qualitative comparison of visual quality across models.

2DGS는 7K iteration에서도 빠른 학습 대비 양호한 품질을 보여주지만, 조명 영역의 빛 번짐이 과도하게 평활화 되어 디테일이 손실되었다. 30K iteration에서는 선명도가 개선되고 주요 패턴이 보다 잘 재현되지만, 여전히 3DGS 및 Mip-Splatting 대비 정밀도가 떨어졌다.

그림 9는 Bonsai 데이터셋에 대한 다양한 모델에 따른 정성적 비교를 제시한다.

NeRF는 전반적으로 blur가 심하고 배경과 객체의 경계가 뭉개져 있으며 심한 왜곡과 잔상 현상이 나타났다. 이는 NeRF가 단순 MLP 기반 학습으로는 깊이 변화가 큰 영역과 고주파 정보를 안정적으로 복원하기 어렵다는 점을 보여준다.

Mip-NeRF는 mip representation을 통해 aliasing 현상을 다소 완화했으나, 여전히 배경 영역이 불분명하고 화분과 같은 세밀한 질감은 흐릿하게 표현되었다.

Instant-NGP는 multi-hash encoding을 통해 전경과 배경의 구조를 상대적으로 선명하게 복원하였다. 창문 밖 영역도 NeRF 계열보다 훨씬 안정적으로 표현되었으며 화분의 꽃도 디테일하게 복원되었다. 그러나 일부 세부 질감에서는 과도한 평활화와 시각적 아티팩트가 관찰되었다.

3DGS는 7K iteration에서 이미 NeRF 계열 모델보다 뛰어난 품질을 보였으나, 표면 질감은 여전히 단순화되었다. 30K iteration에서는 책상 주변의 디테일이 GT에 근접하게 복원되었고, 화분의 꽃도 다른 NeRF 계열 모델에 비해 명확하게 표현되었다.

Mip-Splatting은 iteration에 따른 개선이 두드러졌다. 7K iteration에서는 구조가 안정적이지만 전반적으로 평활화된 경향이 나타났다. 30K iteration에서는 다른 모델보다 창문 밖 깊은 영역의 디테일까지 안정적으로 복원하며, 다중 해상도 표현 덕분에 스케일 변화에도 일관된 품질을 보였다. 다만 일부 세부 질감은 다소 평활하게 표현되었다.

마지막으로 2DGS는 7K iteration에서 빠른 학습 속도 대비 일정 수준의 품질을 보였으나, 창문 밖 영역의 깊이 정보를 제대로 복원하지 못해 구조가 단순화되었다. 30K iteration에서는 전반적인 선명도가 개선되었지만, 여전히 3DGS나 Mip-Splatting 대비 세부 기하 구조와 깊이 표현은 부족하였다.

V. 결 론

본 연구에서는 NeRF와 3DGS 및 후속 모델의 장면 표현 방식의 차이가 Novel View Synthesis 성능에 미치는 영향을 비교하였다. 동일한 데이터셋과 실험 환경에서 각 모델을 학습시킨 후, PSNR, SSIM, LPIPS와 같은 시각 품질 지표와 Training Time, Training Peak Memory의 연산 효율성, 그리고 정성적 결과를 종합적으로 비교·분석하였다. 분석 결과, NeRF와 Mip-NeRF는 전반적으로 blur 현상과 낮은 해상도 문제로 인해 복잡한 장면에서 세밀한 구조를 복원하는 데 한계가 있었다. Instant-NGP는 효율적인 표현 학습을 통해 다른 NeRF 계열 모델에 비해 뚜렷한 개선을 보였으나, 세밀한 질감 표현에서는 여전히 부족함이 존재하였다. 반면 3DGS 계열 모델은 상대적으로 적은 iteration에서도 NeRF 기반 모델들에 비해 우수한 성능을 보였으며, 특히 Mip-Splatting이 모든 화질 지표에서 가장 우수한 성능을 기록하였다. 2DGS는 계산 효율성은 뛰어나지만 깊이 표현과 세밀한 질감 복원에서 제약이 있었다. 이러한 비교 결과는 메모리 사용량 증가에도 불구하고,3D 장면 재구성 분야에서는 시각 품질이 더 중요한 요소이기 때문에 3DGS 계열 모델이 NeRF 계열 모델보다 더 실용적인 대안임을 시사한다. 그러나 본 연구자는 촬영 데이터의 초기 포인트 클라우드 품질이 낮은 경우나 장시간 학습이 가능한 상황에서는 NeRF 계열 역시 충분한 경쟁력을 가질 수 있다고 판단한다. 향후 연구에서는 품질과 메모리, 시간 간의 trade-off 곡선을 제시함으로써 응용 환경별 최적 모델 선택을 체계적으로 지원하고자 한다. 또한 고주파 질감 보존, 깊이 정보 표현력 강화, 복잡한 환경에서의 일반화 성능 검증을 통해 모델의 한계를 보완하고, 학습 효율성 최적화를 바탕으로 실시간 응용 가능성을 확대하는 것이 향후 중요한 연구 방향으로 판단한다.

참 고 문 헌 (References)

[1] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, “View synthesis by appearance flow,” in European Conference on Computer Vision, Cham, Switzerland: Springer, pp. 286-301, Sept. 2016. doi: https://doi.org/10.1007/978-3-319-46493-0_18

[2] 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,” Communications of the ACM, Vol.65, No.1, pp.99-106, 2021.
doi: https://doi.org/10.1145/3503250

[3] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, “Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields,” in Proc. IEEE/CVF Int. Conf. on Computer Vision, Montreal, Canada, pp.5855-5864, Oct. 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00581

[4] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” ACM Transactions on Graphics, Vol.41, No.4, pp.102:1-102:15, July 2022.
doi: https://doi.org/10.1145/3528223.3530127

[5] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3D Gaussian splatting for real-time radiance field rendering,” ACM Transactions on Graphics, Vol.42, No.4, pp.139:1-139:19, 2023.
doi: https://doi.org/10.1145/3592433

[6] Z. Yu, A. Chen, B. Huang, T. Sattler, and A. Geiger, “Mip-Splatting: Alias-free 3D Gaussian splatting,” in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Seattle, USA, pp.19447-19456, June 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01950

[7] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, “2D Gaussian splatting for geometrically accurate radiance fields,” in ACM SIGGRAPH 2024 Conference Papers, Denver, USA, pp.1-11, July 2024.
doi: https://doi.org/10.1145/3641519.3657476

[8] J. Korhonen and J. You, “Peak signal-to-noise ratio revisited: Is simple beautiful?,” in Proc. IEEE Int. Workshop on Quality of Multimedia Experience, Klagenfurt, Austria, pp.37-38, July 2012.
doi: https://doi.org/10.1109/QoMEX.2012.6263845

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, Vol.13, No.4, pp.600-612, Apr. 2004.
doi: https://doi.org/10.1109/TIP.2003.819861

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, USA, pp.586-595, June 2018.
doi: https://doi.org/10.1109/CVPR.2018.00068

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

[12] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, “Local light field fusion: Practical view synthesis with prescriptive sampling guidelines,” ACM Transactions on Graphics, Vol.38, No.4, pp.29:1-29:14, July 2019.
doi: https://doi.org/10.1145/3306346.3322980

[13] A. Knapitsch, J. Park, Q. Y. Zhou, and V. Koltun, “Tanks and Temples: Benchmarking large-scale scene reconstruction,” ACM Transactions on Graphics, Vol.36, No.4, pp.78:1-78:13, July 2017.
doi: https://doi.org/10.1145/3072959.3073599

저 자 소 개

우 성 현

- 2022년 ~ 현재 : 동아대학교 컴퓨터공학과 학사과정
- ORCID : https://orcid.org/0009-0001-0016-5898
- 주관심분야 : 영상 처리, 딥러닝, 컴퓨터비전

서 정 일

- 1994년 : 경북대학교 전자공학과(공학사)
- 1996년 : 경북대학교 대학원 전자공학과(공학석사)
- 2005년 : 경북대학교 대학원 전자공학과(공학박사)
- 1998년 ~ 2000년 : LG반도체 주임연구원
- 2010년 ~ 2011년 : 영국 Southampton University, ISVR 방문연구원
- 2000년 ~ 2023년 : 한국전자통신연구원 실감미디어연구실장
- 2023년 ~ 현재 : 동아대학교 컴퓨터AI공학부 부교수
- ORCID : https://orcid.org/0000-0001-5131-0939
- 주관심분야 : 멀티미디어 부호화, 컴퓨터비전, 실감 영상 및 음향, 멀티미디어 표준화