AI 기반 동적 공간영상 생성 기술 동향

Recent Trends in AI-Based Novel View Synthesis for Dynamic Scenes

임한신 (H.S. Lim, hslim@etri.re.kr) 실감미디어연구실 책임연구원
추현곤 (H.-G. Choo, hyongonchoo@etri.re.kr) 실감미디어연구실 책임연구원/실장

ABSTRACT

Estimating the 3D geometric structure of a scene from images and generating accurate 3D models has been a long-standing research topic and remains an active area of study. However, the process of creating accurate 3D models requires relatively complex steps and achieving the desired quality of a 3D model is often challenging. Considering the complexity and difficulty of generating 3D models, technologies that create novel view images from scene images without precise 3D model information have seen consistent development. Recently, advancements in novel view synthesis technologies that generate images from given scene data have accelerated significantly, particularly after demonstrating the potential of combining these methods with AI technologies to produce higher-quality results. Among AI-based novel view synthesis approaches, AI-based dynamic scene reconstruction and rendering have received significant attention. This study explored recent trends in AI-based novel view synthesis for dynamic scenes.

KEYWORDS AI, NeRF, Novel View Synthesis, 3DGS, 4DGS, 동적공간영상

Ⅰ. 서론

주어진 장면에 대한 영상들로부터 장면의 3D 기하구조를 추정하고 정확한 3D 모델을 생성하는 기술은 이전부터 현재까지 활발히 연구되고 있는 분야이다. 하지만 정확한 3D 모델을 만드는 과정은 비교적 복잡한 세부 과정이 필요하고, 원하는 품질의 3D 모델을 만들기도 쉽지 않다. 이러한 3D 모델 생성의 복잡함 및 어려움으로 인해 장면의 영상들로부터 정확한 3D 모델 정보 없이 사용자가 원하는 가상시점 영상을 생성하는 공간영상 생성 기술 또한 꾸준히 개발되어 왔다[1].

특히 최근 주어진 장면에 대한 영상들로부터 가상시점 영상을 생성하는 기술들이 AI 기술 및 GPU 등의 하드웨어 발전과 결합하면서 이전보다 고품질의 가상시점 영상 생성이 가능함을 보여준 이후[2,3], 공간영상 생성 기술에 대한 관심이 급격히 높

* DOI: https://doi.org/10.22648/ETRI.2025.J.400201
* 본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음[25ZC1110, 초실감 입체공간 미디어콘텐츠 원천기술 연구].

아졌다.

이러한 AI 기반 공간영상 생성 기술들 중 현재 많은 연구가 이루어지고 있는 분야 중 하나가 공간영상 기술을 움직이는 객체가 있는 동영상으로 확장한 AI 기반 동적 공간영상 생성 기술이다.

본고에서는 최근의 AI 기반 동적 공간영상 생성 기술 동향에 대해 알아본다. 앞서 언급한 바처럼 AI 기반 동적 공간영상 생성 기술에 대한 연구가 활발히 이루어지고 있고 이에 따라 다양한 방법이 소개되었다. 본고에서는 이들 중 장면별 학습을 하지 않거나 최소화한 Generalizable Neural Rendering 기술과 다시점 영상을 입력하여 정적 장면을 표현하는 기술인 3DGS(3D Gaussian Splatting) 기술을 동적 장면으로 확장한 4DGS(4D Gaussian Splatting) 기술의 최근 동향에 대해 알아본다.

본고의 구성은 다음과 같다. Ⅱ장에서는 대표적인 최근 공간영상 생성 기술들에 대해 알아본다. Ⅲ장과 Ⅳ장에서는 각각 대표적인 Generalizable Neural Rendering과 4DGS 기술들에 대해서 알아보고, Ⅴ장에서는 AI 기반 동적 공간영상 기술 연구에 사용되는 주요 데이터셋 및 품질측정 지표에 대해 알아본다.

Ⅱ. 최근 공간영상 생성 기술

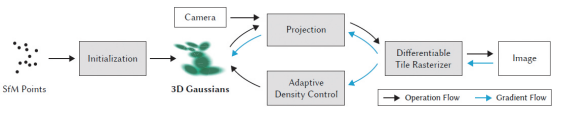
1. NeRF

NeRF(Neural Radiance Fields)의 기본적인 목표는 3D Scene을 Implicit하게 구성하는 MLP(Multi-Layer Perceptron)을 이용하여 가상시점의 영상을 구성하는 것이다[2]. 기본적인 NeRF은 완전 연결 신경망을 사용하여 장면을 각 지점 (x,y,z) 및 방향 (θ,ϕ)을 입력하여 광도(Radiance)와 밀도를 출력하는 연속적인 5D 함수로 표현하며, 이를 카메라 광선(Ray)을 따라 샘플링하여 색상(광도)과 밀도(불투명도)를 예측하도록 학습된다. 또한, 높은 주파수를 표현할 수 있는 Positional Encoding 기법을 적용하여 세밀한 영역에서의 생성 영상의 품질을 높였다. 이후 MLP를 통해 예측된 카메라 광선상의 색상과 밀도로부터 Volume Rendering 기법을 통해 가상시점 영상을 생성하게 된다. 학습 시에는 실제 영상과 생성된 영상 사이의 오차를 미분해 파라미터의 학습을 수행한다.

NeRF는 기존 방법보다 비교적 고품질의 영상 합성이 가능함을 보여줬지만, 일반적으로 긴 학습 시간과 많은 메모리를 필요로 하고 장면별 최적화가 필요하여 고속의 동적 공간영상 생성에는 적용이 쉽지 않다.

2. 3DGS

3DGS은 3D Gaussian들을 Primitive로 하여 장면을 Explicit하게 표현한다[3]. 여기서 각각의 3D Gaussian들은 일반적으로 위치, 회전, 크기, 색상, 불투명도를 파라미터로 가지게 된다. 또한, 3D Gaussian들을 Primitive로 사용함으로써 Differentiable Pipeline이 가능하도록 설계하였다. 렌더링 시에는 타겟 영상의 타일별로 3D Gaussian들을 깊이에 따라 재배열하였고, 이에 기반하여 병렬처리를 통한 효율적인 렌더링이 가능하다. 그림 1은 3DGS의 기본 구조도이다. 하지만 일반적으로 3DGS은 3D Gaussian들의 효율적인 병렬처리를 통해 가상시점 영상 합성은 고속으로 수행할 수 있지만, 각 장면별



출처 Reprinted from B. Kerbl et al., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," ACM Trans. on Graph., no. 4, vol. 42, 2023, pp. 1-14.

그림 1 3D Gaussian Splatting의 기본 구조

로 3D Gaussian들의 파라미터의 최적화 작업 수행이 필요하여 입력 영상으로부터 실시간 생성 등과 같은 고속 처리에 적합하지는 않다. 또한 렌더링 속도면에서 상당한 이점을 제공하고 개별 프레임에서 고품질의 결과를 생성할 수 있지만, 프레임별 방식으로 동적 장면에 적용될 경우에는 메모리, 학습 시간 등의 비효율로 인한 한계를 보인다.

Ⅲ. Generalizable Neural Rendering

1. 기본 개념

앞서 알아본 대표적인 최근 공간영상 생성 기술들인 NeRF와 3DGS의 가상시점 영상의 합성(렌더링)에서는 모든 장면에 대해 Volume Rendering 또는 Gaussian Rasterization이라는 일반화된 렌더링 기법을 적용한다. 하지만 장면의 표현을 위한 네트워크 또는 3D Gaussian들의 파라미터들은 각 장면별로 최적화 과정을 거치게 된다. Generalizable Neural Rendering에서는 이러한 장면별 최적화 과정을 수행하지 않거나 최소화하는 것을 목표로 한다. 입력 영상들로부터 최적화 또는 학습을 위한 반복적인 과정이 생략되므로 영상 입력부터 공간영상 출력까지의 과정을 Feed-Forward의 형태로 구성할 수 있고, 따라서 입력부터 출력까지의 시간을 크게 단축할 수 있다[4-6]. 또한 Stereo Depth Estimation[6-8], MVS(Multi-View Stereo)[9-11] 등의 기법을 적용하여 얻은 사전지식들 또는 중간단계 결과물을 통해 일반화(Generalization) 능력을 높이는 경우가 많다. 또한, 이러한 부가적인 기하 정보들의 사용 및 일반화 능력 때문에 입력 영상들이 비교적 Sparse View에서 동작하는 경우가 많다[5,12,13]. 하지만 장면별 최적화 기법들에 비해 세밀한 영역의 표현력이 떨어진다[14].

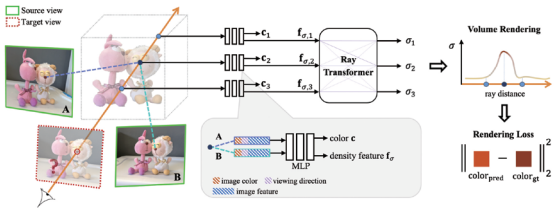
2. 주요 기술들

가. Implicit Representation 기반

1) IBRNet

IBRNet은 NeRF와 같은 Neural Rendering 기법과 전통적인 공간영상 생성 기법인 IBR(Image-Based Rendering)을 결합하여 장면별 학습 없이 일반화(Generalization) 능력을 가지도록 설계되었다[4]. 그림 2는 IBRNet의 기본 구조도이다. IBRNet에서는 각 입력 영상의 특징맵을 추출하고 제안한 IBRNet에서 주어진 광선을 따라 입력 영상들의 정보를 종합 후 타겟 시점에서의 색상을 계산한다. 먼저 타겟 시점 주변의 입력 영상들로부터 CNN 기반의 네트워크를 통해 특징맵을 구한다. 이후 이들로부터 타겟 시점의 각 광선에서의 컬러와 밀도를 MLP와 제안한 Ray Transformer를 통해 구하게 된다. 이후 Volume Rendering을 통해 타겟 시점의 영상을 합성하게 된다.

IBRNet과 같이 입력 영상들의 특징맵들로부터 Volume을 구하고 이들로부터 타겟 시점의 정보를 구하는 방식을 기반으로 하는 다양한 후속 연구들이 있다. 이들 중 참고문헌[9-11,17]은 Volume으로부터 MVS 기법[18]을 적용하였고, 참고문헌[15,16]는 Ray Transformer 기법을 적용하였다.



출처 Reprinted from Q. Wang et al., "IBRNet: Learning Multi-View Image-Based Rendering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021.

그림 2 IBRNet의 기본 구조

2) pixelNeRF

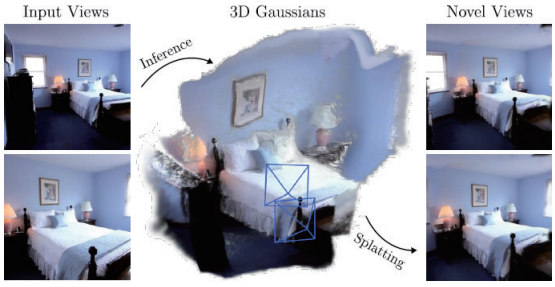
pixelNeRF는 장면들 간의 정보를 재사용하지 못하는 이전의 NeRF 기술들을 개선하기 위해 사전학습된 CNN 기반 부호화기(Encoder)를 통해 생성한 Feature Volume을 타겟 시점으로 투영하여 영상을 생성하였다[12]. 기본적으로 한 장의 영상을 입력으로 생성한 Feature Volume을 타겟 시점으로 투영한 Feature Vector와 위치 및 시점 정보를 NeRF 네트워크의 입력으로 하여 타겟 시점의 Volume Rendering을 위한 정보(색상, 불투명도)를 출력하였다. 또한, pixelNeRF에서의 좌표계는 일반적인 NeRF에서 사용하는 Canonical Space가 아닌 입력 영상이 중심이 되는 Camera Space를 사용하였다.

pixelNeRF는 각 영상마다 생성한 Feature Vector를 NeRF 네트워크의 첫 번째 레이어를 통과 후 이들을 Average Pooling함으로써 다시점 영상까지 확장 가능함을 보여주었다.

나. 3D Gaussian Splatting 기반

1) pixelSplatting

pixelSplatting에서는 기본적으로 두 장의 영상을 입력하여 Two-View Image Encoder와 Pixel-Aligned Gaussian Predictor를 통해 Feed-Forward 방식으로 Gaussian들로 표현된 장면을 생성한다[5]. 그림 3은 pixelSplatting의 개념도이다. 먼저 각각의 영상의 카메라 포즈는 서로 다른 Scale Factor를 가지고 있으므로 두 장의 영상으로 장면을 생성하려면 이러한 Scale Ambiguity 문제를 해결해야 한다. Two-View Image Encoder에서는 Epipolar Transformer[19]를 이용해 영상의 픽셀별 대응점을 찾고 깊이 정보를 추정하여 Scale Ambiguity 문제를 해결하였다. 또한, 3D Gaussian들의 위치를 직접 예측하면 국소 최소값에 빠질 수 있으므로 Pixel-Aligned Gaussian Predictor에서는 깊이에 대한 확률 분포를 예측하고 샘



출처 Reprinted from D. Charatan et al., "pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

그림 3 pixelSplatting의 개념도

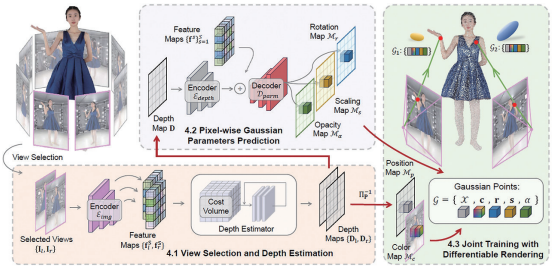
플링하여 Gaussian의 위치를 결정하였다. 이를 통해 pixelSplatting에서는 Scale Ambiguity 문제와 국소 최소값 문제를 해결하고 미분 가능한 학습이 가능하도록 하였다.

현재 본 연구와 관련하여 다양한 후속 연구들이 진행되고 있다[20-24]. 이들 중 MVSplat[20]은 MVS에서 많이 사용하는 기법인 Cost Volume을 적용하여 pixelSplatting에 비해 영상 간 기하구조의 안정성을 높혔다.

2) GPS-Gaussian

GPS-Gaussian에서는 학습 데이터를 이용하여 Depth 정보 추정 네트워크와 3D Gaussian들의 파라미터를 추정하는 네트워크를 사전 학습시킨다[6]. 추론 시에는 사전 학습된 네트워크를 통해 Feed-Forward 방식으로 Gaussian 파라미터를 추정 및 영상 합성을 수행한다.

Depth 정보 추정 시에는 다시점 영상에서 생성하고자 하는 시점 주변의 두 장의 영상을 선택하고 RAFT-Stereo[25]와 같은 스테레오 깊이 추정 방법을 적용하여 두 장의 영상의 깊이맵을 구한다. 이후 스테레오 깊이 추정 부분에서 구한 깊이맵 및 네트워크를 통해 얻어진 영상들의 특징맵들을 입력하여



출처 Reprinted from S. Zheng et al., "GPS-Gaussian: Generalizable Pixel-wise 3D Gaussian Splatting for Real-time Human Novel View Synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

그림 4 GPS-Gaussian의 기본 구조

Gaussian 파라미터 추정부에서 나머지 파라미터들인 회전, 스케일, 불투명도를 구하게 된다. 그리고 이렇게 구해진 파라미터들을 입력하여 각 프레임에서의 가상시점 영상을 생성하게 된다. 이를 통해 RTX 3090에서 약 25fps로 2K의 공간영상 생성까지 가능한 것으로 보고하였다. 그림 4는 GPS-Gaussian의 기본 구조도이다.

이후 본 기술과 관련하여 후속 연구들이 계속 진행되고 있다[26-28]. 이들 중 참고문헌[26]은 특징맵들의 Splatting을 적용하여 타겟 시점의 특징맵을 생성하고 이를 영상으로 변환하였고, 참고문헌[27]은 가상시점 영상의 생성 범위를 배경까지 확장하였다.

Ⅳ. 4D Gaussian Splatting

1. 기본 개념

4DGS는 3DGS 기술을 동적 영상으로 확장한 기술이다. 정적 장면의 표현을 기본으로 하는 3DGS을 바탕으로 시간의 변화에 따른 Gaussian들의 변화를 정확하고 효율적으로 표현하는 것을 목표로 하고 있고 이를 위한 다양한 방법들이 나오고 있다. 4DGS를 분류하는 여러 기준 중 본고에서는 3D Gaussian들의 변화를 표현 및 업데이트하는 방식에 따라 Deformation 기반[29-31], Spatio-Temporal 기반[32-34], Per-Fframe Training[35-37] 기반으로 나누고 각 방식의 주요 기술들에 대해 소개한다.

2. 주요 기술들

가. Deformation 기반 4DGS

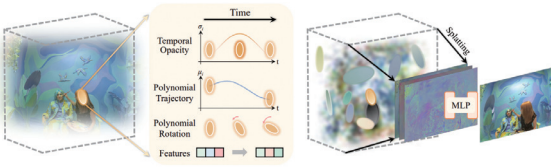
Deformation 기반 4DGS에서는 Gaussian들의 파라미터들의 시간 변화에 따른 Deformation 값(Δ파라미터)을 추정하고 이를 시간에 따라 업데이트한다(파라미터=파라미터(t-α)+Δ파라미터)[29-31].

Deformation 기반 4DGS 기술들 중 하나인 4D Gaussian Splatting[29]은 6개의 3D 평면 조합으로 동적 장면을 표현하는 HexPlanes[38] 기술을 3D Gaussian에 적용하여 Deformation을 추정하였다. 먼저 부호화단에서는 Gaussian의 공간적 위치와 타임스탬프가 6개의 3D 평면에 투영되어 평면 특징맵을 생성한다. 이때 특정 해상도의 경계 사각형 내의 네 모서리에 있는 특징 벡터를 보간하여 해당 해상도에서 투영된 Gaussian의 중심의 특징값을 계산하고 해상도를 증가시켜 또 다른 특징값을 얻는다. 이러한 다양한 해상도에서의 특징값들을 결합(Concatenation)하고, 이를 MLP에 전달하여 최종 특징벡터를 구한다.

부호화단에서 구한 특징벡터들은 세 개의 MLP로 전달된다. 이 MLP들의 출력으로 원래의 3D Gaussian의 위치, 회전, 크기 파라미터들의 Deformation 값들이 나오고 이를 원래의 Gaussian과 결합하여 업데이트를 수행한다.

나. Spatio-Temporal 기반 4DGS

Spatio-Temporal 기반 4DGS에서는 Gaussian들의 파라미터들이 시간이라는 매개변수를 통한 함수의



(a) (b)

출처 Reprinted from Z. Li et al., "Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

그림 5 GPS-Gaussian의 기본 구조 Spacetime Gaussian의 개념도: (a) Spacetime Gaussians, (b) Feature Splatting and Rendering

값으로 표현된다(파라미터=파라미터(t))[32-34].

Spatio-Temporal 기반 4DGS 기술들 중 하나인 Spacetime Gaussian[33]에서는 각 Gaussian의 시간적 중심과 현재 프레임의 타임스탬프 간의 차이를 매개변수로 하여 Gaussian들의 위치와 회전, 불투명도 파라미터들을 함수로 표현하였다. 그림 5는 Spacetime Gaussian의 개념도이다.

구체적으로 Spacetime Gaussian은 동적 장면에서 부드러운 전환을 위해 다항식을 사용하여 위치와 회전을 모델링하였다. 또한 Gaussian의 불투명도(Opacity)는 시간적 중심을 중심으로 하는 1D Gaussian으로 모델링하여 시간에 따라 값이 변하도록 하였다. 색상의 경우 메모리 비효율적인 Spherical Harmonics 대신 기본 색상과 시점과 시간의 정보를 갖고 있는 특징들로 부호화하고 MLP로 이들을 복호화하거나(Full 버전) 기본 색상만을 복호화하여 가상시점 영상을 생성하였다(Lite 버전).

다. Per-Frame Training 기반 4DGS

Deformation 기반 및 Spatio-Temporal 기반의 방법들은 일반적으로 전체 또는 구간별 프레임들을 입력하여 최적화를 수행한다. 반면 Per-Frame Training 기반 4DGS에서는 각 프레임별로 이전 프레임들에서의 Gaussian들의 파라미터들로부터 현재 프레임에서의 파라미터들을 추정 및 업데이트한다[35-37].

Per-Frame Training 기반 4DGS 방법들 중 하나인 3DGStream[35]은 장면의 전체 프레임에 대해 학습하는 대신 프레임별 학습을 통해 동적 공간영상을 생성하도록 설계되었으며, 공간영상을 생성하면서 온라인 스트리밍의 가능성을 보여주는 것을 목표로 하였다.

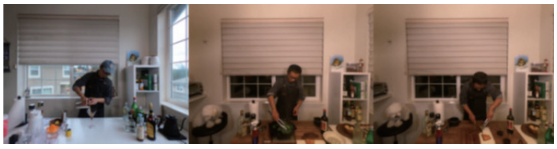
3DGStream은 크게 두 단계의 파이프라인으로 이루어졌다. 첫 번째 단계에서는 Neural Transformation Cache라고 불리는 메모리 및 속도 효율적인 MLP를 사용하여 이전 프레임에서의 Gaussian들의 변화를 예측하고 이를 현재 프레임의 Gaussian에 추가한다. 두 번째 단계에서는 현재 프레임의 Gaussian 위에 새로운 Gaussian을 생성하여 현재 프레임에서 새롭게 나타나는 객체를 효율적으로 표현하도록 하였다. 두 단계의 파이프라인이 완료된 후 현재 프레임에서의 공간영상이 생성되며 첫 번째 단계의 Gaussian만이 다음 프레임으로 전달된다.

Ⅴ. 데이터셋 및 품질측정

1. 데이터셋

AI 기반 동적 공간영상 기술은 기반이 되는 AI 기술의 핵심 업무 및 적용 범위에 따라 학습 및 테스트에 사용하는 데이터셋의 종류가 달라진다. 일반적으로 잘 알려진 Nerf synthetic, LLFF[39] 데이터셋 외에 Generalizable Neural Rendering에서 쓰이는 대표적인 데이터셋으로는 THuman 2.0, 2.1(사람)[40,41], RealEstate10K(공간)[42], DTU[43], ZJU-MoCap[44] 등이 있고, 4DGS에서 주로 많이 쓰이는 데이터셋은 DyNeRF[45], HyperNeRF[46], D-NeRF[47], ENeRF[48] 등이 있다. 카메라 파라미

(a)



(b)

출처 Reprinted from T. Zhou et al., "Stereo magnification: Learning view synthesis using multiplane images," ACM Trans. Graph., vol. 37, no. 4, 2018, pp. 1–12.

Reprinted from T. Li et al., "Neural 3d video synthesis from multi-view video," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (New Orleans, LA, USA), Jun. 2022.

그림 6 4DGS에서 쓰이는 대표적인 데이터셋들: (a) RealEstate10K 데이터셋, (b) DyNeRF 데이터셋

터의 경우 데이터셋이 실제영상인 경우 보통 Colmap[49] 등의 카메라 파라미터 추정 SW를 통해 구한 값을 제공한다. 그림 6은 각각 (a) RealEstate10K, (b) DyNeRF 데이터셋 영상들이다.

가. THuman 2.0, 2.1 데이터셋

THuman2.0 데이터셋은 인간의 3D 재구성 및 표현을 연구하기 위해 만들어진 3D 인간 데이터셋으로 칭화대학교에서 제작하고 배포하였다[40,41]. 이 데이터셋은 현재 컴퓨터 비전, 그래픽스, 및 머신 러닝 연구자들이 인간의 3D 모델링, 포즈 추정 및 시각적 재구성에 가장 많이 활용하고 있는 데이터셋 중 하나이다.

일반적으로 THuman2.0 데이터셋은 총 512개의 3D 모델 중 100개의 모델이 검정(Validation)에 사용되고 나머지는 학습에 사용된다. 2024년에 모델의 수가 약 2,500개로 증가된 THuman 2.1 데이터셋이 배포되었다.

나. RealEstate10K 데이터셋

RealEstate10K는 Youtube에 업로드된 건축물의 실내외 비디오들로 만든 대규모의 데이터셋이다[42]. 약 10,000개의 비디오들로부터 각각 수십에서 수백 프레임의 크기를 갖는 80,000여 개의 클립을 제작하였다. 총 프레임의 수는 약 10,000,000여 장 정도이고, 해상도는 1,920×1,080이다. 또한, 각각의 클립에 대해 SLAM과 Bundle Adjustment 알고리즘을 이용하여 각 프레임마다의 카메라 파라미터들을 구하여서 제공하고 있다. 전체 클립들 중에서 약 90% 정도가 학습 데이터셋이고, 10% 정도가 테스트용 데이터셋이다.

다. DyNeRF 데이터셋

DyNeRF 데이터셋은 4DGS 포함 가상시점 영상 합성 기술의 성능 평가에 널리 사용되는 데이터셋이다[45]. 이 데이터셋은 다양한 주방 작업을 수행하는 한 사람을 중심으로 하는 여섯 개의 실내 장면으로 구성되어 있다. DyNeRF 데이터셋은 20개의 서로 다른 카메라로 다양한 각도에서 촬영된 10~40초 길이의 동영상으로 구성되어 있고 각 동영상은 30fps로 되어 있다. 4D Gaussian Splatting에 관한 대부분의 연구는 각 카메라 각도에 대해 처음 300프레임의 결과를 측정한다. 동영상은 2,704×2,028 해상도로 제공되며, 대부분의 연구에서는 이를 1,352×1,014로 다운샘플링하여 사용한다.

2. 품질측정 지표

동적 공간영상의 정량적인 품질측정은 일반적으로 각 시간이나 프레임에서의 원본 영상과 원본 영상과 동일한 시점, 좌표 및 내부 카메라 파라미터 환경에서 생성된 가상시점 영상과의 유사 정도를 측정함으로써 이루어진다. 대표적인 지표들로는

PSNR, SSIM[50], LPIPS[51] 등이 있다.

가. PSNR

PSNR(Peak Signal-to-Noise Ratio)은 재구성된 이미지와 참조 이미지의 품질을 측정하는 데 사용되는 가장 일반적인 지표이다. 측정단위는 dB이며 두 영상 간의 각 픽셀값의 차이의 제곱을 기반으로 한다. PSNR 값이 높을수록 원본 영상과 생성된 영상이 서로 유사하다는 것을 의미한다.

나. SSIM

SSIM(Structural Similarity Index Measure)은 영상의 구조적 정보, 밝기, 대비의 변화를 고려하여 품질을 측정한다[50]. 이를 통해 SSIM은 인간의 인식과 더 유사하게 두 영상의 구조가 얼마나 유사한지를 평가한다. SSIM 값은 -1에서 1 사이로, 1은 완전히 같은 영상임을 나타내고 0은 유사성이 없음을 -1은 완전한 반상관임을 나타낸다.

다. LPIPS

LPIPS(Learned Perceptual Image Patch Similarity)는 VGG, AlexNet, SqueezeNet 등의 CNN 기반 딥러닝 네트워크를 통해 추출된 특징을 기반으로 두 영상 간의 유사성을 측정하는 인지적 지표이다[51]. 이러한 특징의 유사성의 측정은 인간의 시각적 인식과 상관관계가 높은 것으로 알려져 있다. LPIPS 값의 범위는 0에서 1 사이이고, 낮은 LPIPS 값은 영상들이 인지적으로 더 유사하다는 것을 의미한다.

VI. 결론

본고에서는 최근의 AI 기반 동적 공간영상 생성 기술 동향에 대해 알아보았다. 구체적으로 장면별 학습을 하지 않거나 최소화한 Generalizable Neural Rendering 기술과 3DGS 기술을 동적 객체가 있는 동영상으로 확장한 4DGS 기술의 최근 동향에 대해 알아보았다. 현재 본 분야는 기술의 트렌드 변화가 매우 빠르고 지금도 관련하여 새로운 기술들이 쏟아지고 있는 분야이다. 이러한 본 분야의 최신 기술 흐름에 뒤처지지 않기 위해서는 관련 기관들과의 정보 교류 및 협력이 매우 중요하다고 하겠다.

용어해설

Volume Rendering 색상 및 불투명도 등의 정보를 가지고 있는 복셀에 광선을 투과하고 광선상의 색상을 불투명도에 따라 누적함으로써 타겟 시점에서의 영상을 생성하는 방법

Primitive 장면을 구성하고 표현 및 처리할 수 있는 가장 기본적인 기하학적 구조

IBR Image-Based Rendering의 약자. 명시적인 3D 모델이 없는 상태에서 입력 영상들로부터 새로운 시점의 영상을 생성하는 기술

약어 정리

3DGS	3D Gaussian Splatting
LPIPS	Learned Perceptual Image Patch Similarity
NeRF	Neural Radiance Field
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure

참고문헌

[1] H. Shum and S.B. Kang, “Review of image-based rendering techniques,” in Proc. Vis. Commun. and Image Process., 2000, pp. 2-13.

[2] B. Mildenhall et al., “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” in Proc. Eur. Conf. Comput. Vis., Aug. 2020.

[3] B. Kerbl et al., “3D Gaussian Splatting for Real-Time Radiance Field Rendering,” ACM Trans. Graph., vol. 42, no. 4, 2023, pp. 1-14.

[4] Q. Wang et al., “IBRNet: Learning Multi-View Image-Based Rendering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021.

[5] D. Charatan et al., “pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D

Reconstruction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[6] S. Zheng et al., "GPS-Gaussian: Generalizable Pixel-wise 3D Gaussian Splatting for Real-time Human Novel View Synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[7] Y. Liu et al., "Neural Rays for Occlusion-aware Image-based Rendering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (New Orleans, LA, USA), Jun. 2022.

[8] H.C. Lee et al., "Generalizable Novel-View Synthesis using a Stereo Camera," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[9] A. Chen et al., "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in Proc. IEEE/CVF Int. Conf. Comput. Vis., Jun. 2021.

[10] M. Johari et al., "GeoNeRF: Generalizing NeRF with Geometry Priors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (New Orleans, LA, USA), Jun. 2022.

[11] T. Liu et al., "MVSGaussian: Fast Generalizable Gaussian Splatting Reconstruction from Multi-View Stereo," in Proc. Eur. Conf. Comput. Vis., (Milano, Italy), Sep. 2024.

[12] A. Yu et al., "pixelNeRF: Neural Radiance Fields from One or Few Images," in Proc. IEEE/CVF Int. Conf. Comput. Vis., Jun. 2021.

[13] S. Szymanowicz et al., "Splatter Image: Ultra-Fast Single-View 3D Reconstruction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[14] Y.H. Yoon et al., "GMT: Enhancing Generalizable Neural Rendering via Geometry-Driven Multi-Reference Texture Transfer," in Proc. Eur. Conf. Comput. Vis., (Milano, Italy), Sep. 2024.

[15] M. Suhail et al., "Generalizable patch-based neural rendering," in Proc. Eur. Conf. Comput. Vis., (Tel Aviv, Israel), Oct. 2022.

[16] M. Varma et al., "Is Attention All That NeRF Needs?," Int. Conf. Learn. Representations., (Kigali, Rwanda), May. 2023.

[17] H. Xu et al., "MuRF: Multi-Baseline Radiance Fields," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[18] Y. Yao et al., "MVSNet: Depth Inference for Unstructured Multi-view Stereo," in Proc. Eur. Conf. Comput. Vis., (Munich, Germany), Sep. 2018.

[19] Y. He et al., "Epipolar transformers," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020.

[20] Y. Chen et al., "MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-View Images," in Proc. Eur. Conf. Comput. Vis., (Milano, Italy), Sep. 2024.

[21] C. Wewer et al., "latentSplat: Autoencoding Variational Gaussians for Fast Generalizable 3D Reconstruction," in Proc. Eur. Conf. Comput. Vis., (Milano, Italy), Sep. 2024.

[22] Y. Chen et al., "MVSplat360: Feed-Forward 360 Scene Synthesis from Sparse Views," in Proc. Adv. Neural Inf. Process. Syst., Dec. 2024.

[23] Y. Wang et al., "FreeSplat: Generalizable 3D Gaussian Splatting Towards Free-View Synthesis of Indoor Scenes," in Proc. Adv. Neural Inf. Process. Syst., Dec. 2024.

[24] H. Xu et al., "DepthSplat: Connecting Gaussian Splatting and Depth," arXiv preprint, 2024. doi: 10.48550/arXiv.2411.04924

[25] L. Lipson et al., "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching," in Proc. Int. Conf. 3D Vis., Dec. 2021.

[26] Hanzhang Tu et al., "Tele-Aloha: A Telepresence System with Low-budget and High-authenticity Using Sparse RGB Cameras," in Proc. ACM SIGGRAPH Conf., (Denver, CO, USA), Aug. 2024.

[27] B. Zhou et al., "GPS-Gaussian+: Generalizable Pixel-wise 3D Gaussian Splatting for Real-Time Human-Scene Rendering from Sparse Views," arXiv preprint, 2024. doi: 10.48550/arXiv.2411.11363

[28] Y. Hu et al., "EVA-Gaussian: 3D Gaussian-based Real-time Human Novel View Synthesis under Diverse Camera Settings," arXiv preprint, 2024. doi: 10.48550/arXiv.2410.01425

[29] G. Wu et al., "4D gaussian splatting for real-time dynamic scene rendering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[30] Z. Yang et al., "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[31] J. Bae et al., "Per-Gaussian Embedding-Based Deformation for Deformable 3D Gaussian Splatting," in Proc. Eur. Conf. Comput. Vis., (Milano, Italy), Sep. 2024.

[32] Z. Yang et al., "Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting," Int. Conf. Learn. Representations., (Vienna, Austria), May. 2024.

[33] Z. Li et al., "Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[34] Y. Lin et al., "Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle," in Proc. IEEE/CVF Conf.

Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[35] Jiakai Sun et al., "3DGStream: On-the-fly Training of 3D Gaussians for Efficient Streaming of Photo-Realistic Free-Viewpoint Videos," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), Jun. 2024.

[36] J. Lee et al., "Fully Explicit Dynamic Gaussian Splatting," in Proc. Adv. Neural Inf. Process. Syst., Dec. 2024.

[37] Z. Liu et al., "Dynamics-Aware Gaussian Splatting Streaming Towards Fast On-the-Fly Training for 4D Reconstruction," arXiv preprint, 2024. doi: 10.48550/arXiv.2411.14847

[38] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Vancouver, Canada), Jun. 2023.

[39] B. Mildenhall et al., "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," ACM Trans. Graph, vol. 38, 2019, no. 4, pp. 1–14.

[40] T. Yu et al., "Function4d: Real-time human volumetric capture from very sparse consumer RGBD sensors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021.

[41] https://github.com/ytrock/THuman2.0-Dataset

[42] T. Zhou et al., "Stereo magnification: Learning view synthesis using multiplane images," ACM Trans. Graph., vol. 37, no. 4, 2018, pp. 1–12.

[43] R. Jensen et al., "Large Scale Multi-view Stereopsis Evaluation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (Columbus, OH, USA), Jun. 2014.

[44] S. Peng et al., "Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021.

[45] T. Li et al., "Neural 3d video synthesis from multi-view video," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (New Orleans, LA, USA), Jun. 2022.

[46] K. Park et al., "Hypernerf: A higherdimensional representation for topologically varying neural radiance fields," ACM Trans. Graph., vol. 40, no. 6, 2021, pp. 1–12.

[47] A. Pumarola et al., "D-nerf: Neural radiance fields for dynamic scenes," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021.

[48] H. Lin et al., "Efficient Neural Radiance Fields for Interactive Free-viewpoint Video," in Proc. SIGGRAPH Asia Conf., (Daegu, Korea), Dec. 2022.

[49] J.L. Schonberger and J.M. Frahm, "Structure-from-Motion Revisited," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Las Vegas, NV, USA), Jun. 2016.

[50] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, 2004, pp. 600–612.

[51] R. Zhang et al., "Lpips: A learned perceptual image patch similarity metric," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Salt Lake City, UT, USA), Jun. 2018.