6자유도 몰입형 비디오 합성을 위한 동적 방사형 필드 모델링 기법

박준형 류은석
성균관대학교 실감미디어공학과
{joke0702, esryu}@skku.edu

Dynamic Radiance Fields Modeling for 6-DoF Immersive Video Synthesis

Jun-Hyeong Park Eun-Seok Ryu
Department of Immersive Media Engineering, Sungkyunkwan University

요 약

본 논문은 Neural Radiance Fields (NeRF)의 신경망 기반 방사형 필드 표현을 확장하여 6 자유도 가상현실에서 몰입형 비디오 합성을 위한 접근법을 제안한다. 복셀 그리드를 활용한 효율적인 방사형 필드 표현을 통해 NeRF의 계산적 오버헤드를 줄이고, 복잡한 실사 데이터를 높은 품질로 복원했다. 더 나아가, 긴 프레임의 동적 방사형 필드를 표현하기 위해 프레임 간의 유사정보를 활용하여 필수 정보만을 추출하고 모델링하는 기법을 도입하였다. 실험적 검증을 통해, 제안된 방법론이 프레임 별 독립적인 학습을 진행하는 기존 동적 방사형 필드 표현 기술 대비 시간 절감과 자원의 효율적 사용을 가능케 하며, 동적 콘텐츠의 일관성을 보장하면서도 우수한 복원 품질을 제공함을 입증하였다. 이러한 결과는 가상현실 내 신경망 기반 비디오 합성과 관련된 연구 및 응용 분야의 새로운 가능성을 제시한다.

1. 서론

6 Degrees of Freedom (6DoF)을 지원하는 몰입형 비디오 기술은 사용자가 가상공간을 자유롭게 이동하며 어떠한 각도와 위치에서든 콘텐츠를 관찰할 수 있는 새로운 차원의 경험을 가능하게 한다. 이러한 기술은 head-mounted display (HMD)를 통해 가상현실 체험을 하는 사용자에게 실제 환경과 유사한 몰입감과 생동감을 제공한다. 엔터테인먼트, 교육, 관광, 전문 시뮬레이션 등 다양한 분야에서 몰입형 비디오 기술은 혁신의 가능성을 내포하고 있으며, 기술의 발전이 해당 콘텐츠의 생성 및 소비 장벽을 낮춤에 따라 실감형 미디어 시장은 크게 확장될 것으로 예상된다.

그러나 이러한 경험을 구현하는 과정에는 컨텐츠의 생성에서부터 전송, 그리고 최종적인 소비에 이르기까지 다양한 기술적 도전 과제들이 존재한다. 고품질의 몰입형 영상 콘텐츠 생성을 위해서는 모든 방향에서 고해상도 이미지를 캡처할 수 있는 첨단 카메라 시스템과, 이들 카메라 간의 정밀한 동기화가 요구된다. 또한, 캡처 된 이미지들을 통합하여 하나의 영상으로 합성하는 과정은 방대한 데이터의 효율적 관리와 높은 처리 능력을 필요로 한다. 현존하는 비디오 코덱 기술은 이러한 고사양 요구 사항을 충족하는 데 한계를 지니고 있다. 스트리밍 단계에서는 고대역폭 요구와 사용자 상호작용에 기반한 동적인 스트리밍 품질의 조정이라는 두 가지 복잡한 문제에 직면한다. 이는 몰입형 미디어를 원활하게 전송하는 것뿐만 아니라, 사용자 경험의 품질을 지속적으로 유지하고 개선하는 데 필수적인 요소들이다. 이러한 기술적 한계를 극복하기 위해, 몰입형 영상의 제작부터 배포, 그리고 재생에 이르는 전 과정에 걸쳐 조화를 이루는 표준화된 기술 프레임워크의 개발이 요구되고 있으며, 이에 대한 표준화 작업이 국제 표준화 기구인 moving picture experts group (MPEG) working group (WG) 4에서 활발하게 진행 중이다[1].

몰입형 비디오의 생성을 위한 접근법 중 하나는 기하학 기반 방식이다. 이는 3차원 메쉬나 포인트 클라우드를 재구성하여 실세계의 장면을 디지털화 하는 방식이다. 이 접근법은 상세한

장면의 정밀한 디테일을 재현하는 데 있어서 일정한 성공을 거두었으나, 물체의 가려짐 현상이나 질감이 없는 영역의 정확한 정보를 추출하는 데 한계를 보였다. 또한, 실제와 같은 조명과 그림자 효과를 모델링하는 것은 매우 복잡한 작업으로, 전통적인 기하학적 방법론으로 만족스러운 결과를 얻기 어려웠다.

또 다른 접근법인 Multi-Plane Images (MPI) 방식은 여러 평면 이미지를 사용하여 깊이와 관점에 따른 변화를 나타내는 방법이다. 이 방식은 각 평면에 캡처 된 이미지를 배치하여 시점 변화에 따라 다른 이미지를 보여준다. 그러나 MPI는 시점에 따라 결과가 크게 달라질 수 있는 종속성, 높은 메모리 요구량, 그리고 고정된 해상도의 한계로 인해 복잡한 시각적 세부사항을 재현하는 데 제약이 있었다. 특히, 시점이 크게 변경될 때 비현실적인 결과를 초래할 수 있고, 고해상도를 처리하기 위한 상당한 계산 자원을 필요로 했다.

최근에는 NeRF[2]와 같은 신경망 기반 접근법이 제안되었다. NeRF는 딥러닝을 활용해 3차원 장면의 복잡한 라이팅과 재질 특성을 효과적으로 학습한다. 이를 통해, 다양한 시점에서 일관된 고품질 렌더링을 생성할 수 있으며, 이는 기존 방법들이 어려움을 겪었던 다양한 조명 조건과 복잡한 재질 효과를 사실적으로 재현했다. 그러나 NeRF는 정적인 장면에 초점을 맞춰 개발되었으며, 동적 장면으로 확장하려는 시도는 추가적인 도전과제를 수반한다. 또한 높은 계산 비용으로 인해 실시간 처리가 어렵다는 문제도 여전히 해결해야 할 과제로 남아 있다.

본 논문에서는 기존 NeRF의 한계를 극복하고자 최신 SOTA 모델을 발전시켜, 다각도에서 촬영된 다수의 동영상과 카메라 메타데이터를 통해 6DoF 몰입형 비디오를 효율적으로 생성하는 시스템 프레임워크를 제시한다. 본 시스템은 3차원 복셀 그리드 기반의 외재적인 데이터 표현을 사용해 NeRF의 렌더링 속도를 개선하고 프레임간 유사 정보를 활용하여 긴 프레임의 동적 장면을 효율적으로 모델링한다.

2. 선행 연구

NeRF는 3차원 장면을 재구성하기 위해 방사형 필드의 색상과 밀도를 Ψ(x,d)로 표현하는데, 여기서 x는 3차원 공간 좌표, d는 2차원 시야 방향을 나타낸다. 이러한 5차원의 입력 벡터를 MLP 디코더를 사용하여 방사형 필드로 근사하고, 볼륨 렌더링을 적용하여 임의의 시점에서의 장면을 재구성한다. 그러나 NeRF의 훈련과 추론 과정은 많은 연산량을 요구해 실시간 응용에 많은 제약이 따른다.

선행 연구인 Direct Voxel Grid Optimization (DVGO)[3]는 이러한 문제를 해결하고자 명시적 그리드 표현 방식을 채택했다.

DVGO는 정적 장면의 방사형 필드를 명시적인 밀도 그리드 (Vσ) 와 색상 피처 그리드 (Vc)를 활용하여 모델링한다:

$$\sigma = \text{interp}(\mathbf{x}, \mathbf{V}_\sigma)$$
$$\mathbf{c} = \Phi(\text{interp}(\mathbf{x}, \mathbf{V}_\mathbf{c}), \mathbf{d})$$

여기서 interp(·)는 삼선형 보간(trilinear interpolation)을 나타내며, 이는 특정 위치에서 밀도 값과 색상 값을 그리드 상의 값들을 보간하여 계산하는 기법이다. 특정 시야 방향에서의 최종 색상 값은 Vc를 보간한 후 추가적으로 얕은 MLP(Φ)를 통해 결정된다. 추후 연산의 복잡성을 줄이기 위해, Vσ와 Vc는 통합된 단일 피처 그리드로 병합되어 사용된다. 이러한 외재적 데이터와 내재적 네트워크를 혼합하여 방사형 필드를 표현하는 접근법은 기존 NeRF의 복원 품질을 유지하면서도 연산 효율성을 개선하는 장점을 가진다.

DVGO가 정적인 방사형 필드를 성공적으로 복원했지만, 동적 장면을 처리하기 위한 기술 확장이 요구되었다. Residual Radiance Fields (ReRF)[4]는 인접한 프레임 간의 잔차 정보를 명시적으로 모델링하여 동적 장면의 복잡한 움직임과 변화를 효과적으로 학습하는 새로운 방법론을 제안한다.

ReRF는 DVGO의 명시적 그리드 표현을 동적 장면에 적용했다. 첫 번째 프레임을 온전한 피처 그리드로 표현하고, 후속 프레임들을 모션 그리드와 잔차 그리드를 활용하여 나타낸다. 모션 그리드는 저해상도의 그리드를 사용하여 프레임 간의 상대적 움직임을 나타내고, 잔차 그리드는 인접 프레임 간의 차이와 새로운 정보를 나타낸다. DVGO와 유사하게, ReRF는 공간 좌표와 시야 방향을 입력으로 사용하고, 특정 위치와 시간에서의 방사형 필드 값을 출력하는 글로벌 MLP를 디코더로 활용한다.

이러한 구조 덕분에, ReRF는 기존의 동적 NeRF 기법들이 의존하던 표준 피처 공간을 사용하지 않아 큰 모션과 토폴로지 변화가 있는 장면들을 안정적으로 재구성할 수 있으며, 긴 시퀀스의 동적 장면을 효과적으로 복원한다. 그러나 ReRF의 적용 범위는 객체 데이터에 국한되어, 실사 장면 데이터의 복원으로 확장하기 위해서는 추가적인 연구가 필요하다.

3. 본론

본 논문에서 제안하는 시스템은 선행 연구인 ReRF 모델을 기반으로 실사 장면 데이터의 복원을 수행한다. 기존 ReRF 모델을 사용하여 장면 데이터를 학습하고 복원을 시도하였을 때 객체 중심 데이터에 최적화된 기존 모델의 한계로 인해 불분명한 형태의 결과물이 렌더링 되었다.

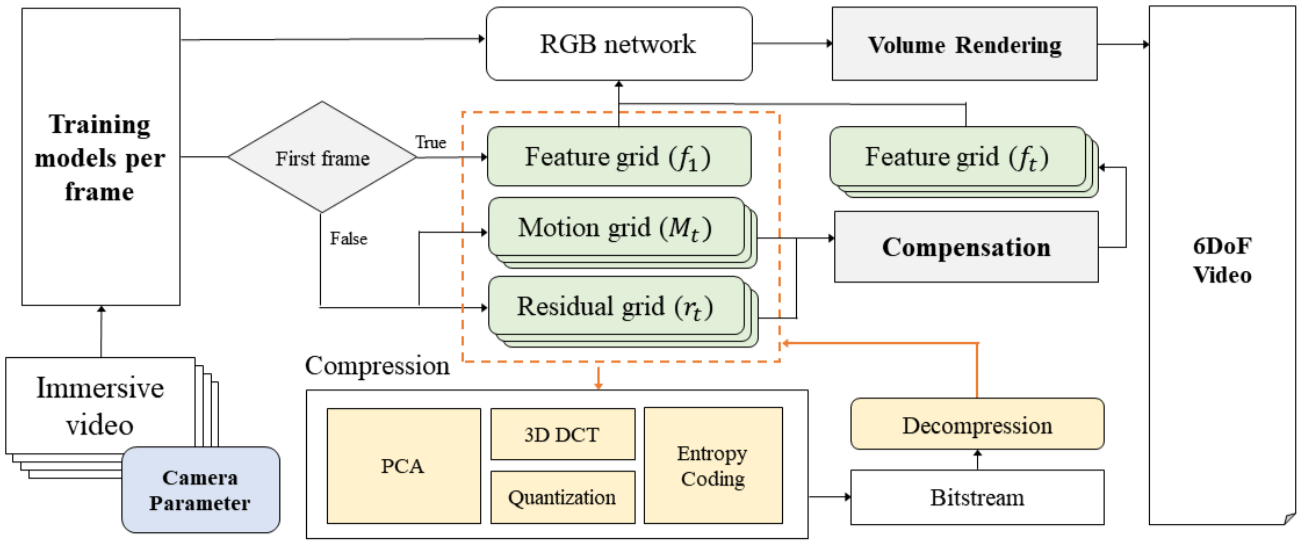

사진 1. 복셀 그리드 기반 6 자유도 영상 합성 시스템 구조도

이 문제를 해결하기 위해, 실사 데이터의 특성을 적절히 처리할 수 있는 방법으로 선행 연구인 Local Light Field Fusion (LLFF)[5]의 접근법을 차용하였다. LLFF에서 실제 장면의 기하학적 및 광학적 일관성을 보존하기 위해 정규화 디스플레이 좌표(NDC) 변환을 적용한다. NDC는 3차원 공간의 포인트를 2차원 화면 공간으로 매핑할 때 발생하는 왜곡을 최소화하며, 멀티뷰 렌더링에서 각각의 뷰포인트에 대해 일관된 깊이 정보를 유지하는 데 중요한 역할을 한다. 이러한 접근을 통해, 본 시스템은 각 프레임에 걸쳐 변화하는 장면의 깊이를 효과적으로 처리하고, 다양한 시점에서 복원된 장면의 시각적 일관성을 개선한다. ReRF의 데이터 로딩 과정에서 NHR 포맷 대신 LLFF 포맷을 사용하도록 시스템의 구성을 변경하고, NDC변환을 적용하여 데이터의 기하학적 및 광학적 일관성을 유지하였다.

사진 1. 은 본 논문에서 제안하는 시스템의 전체 구조도를 나타낸다. 이러한 접근법을 통해 기존 모델 대비 향상된 품질로 장면 데이터를 복원할 수 있었으며, 복원된 장면은 객관적인 품질 지표뿐만 아니라 다양한 시점에서 시각적 일관성 면에서도 우수한 결과를 보여주었다. 추가로 더 나은 결과를 달성하기 위해 하이퍼파라미터 최적화를 실시했다. 이 과정에서 학습 반복 횟수, 복셀 그리드의 해상도, 광선 샘플링 빈도, 그리고 학습률을 조정하여 장면의 고주파 세부 사항을 더욱 잘 복원할 수 있었다. 이런 조정을 통해 장면 데이터의 미세한 디테일을 보존하고 전체적인 품질 지표를 개선했다.

본 연구의 실험 설계는 100프레임으로 구성된 15개의 멀티뷰 비디오 데이터셋을 입력으로 사용한다. 이 데이터셋은 다양한 시점에서 캡처된 비디오 프레임들을 포함하며, 각 비디오는 카메라 메타데이터 정보를 함께 제공한다. 실험은 두 단계의 모델을 사용하여 수행되었다. 첫 번째 단계에서는 NDC 변환을 적용하여 ReRF 모델을 변형한 모델을 사용하였다. 두 번째 단계에서는 추가적인 하이퍼파라미터 최적화를 적용한 모델을 사용해 실험을 진행했다.

학습 과정에서 데이터셋을 훈련용 뷰와 테스트용 뷰로 분리하였다. 훈련용 뷰는 모델의 학습에 사용되며, 테스트용 뷰는 학습 과정에서 사용되지 않아, 모델의 일반화 능력을 검증하는 데 활용된다. 실험의 객관성을 확보하기 위해, 모델의 품질 평가는 훈련에 사용되지 않은 테스트용 뷰에서 첫 프레임을 렌더링한 결과물을 바탕으로 이루어졌다.

사진 2. 은 테스트 뷰에서 렌더링한 결과의 품질을 시각적으로 비교할 수 있도록 나타내었다. 이는 NDC 변환만 적용된 모델과 하이퍼파라미터 최적화가 적용된 모델의 출력 결과를 직접 비교함으로써, 본 연구에서 제안된 방법론의 효과를 명확히 보여준다. 이러한 방식으로 본 연구는 ReRF 모델의 확장과 최적화를 통해 복셀 그리드 기반 방식으로 동적 장면의 정밀한 복원이 가능함을 입증한다.

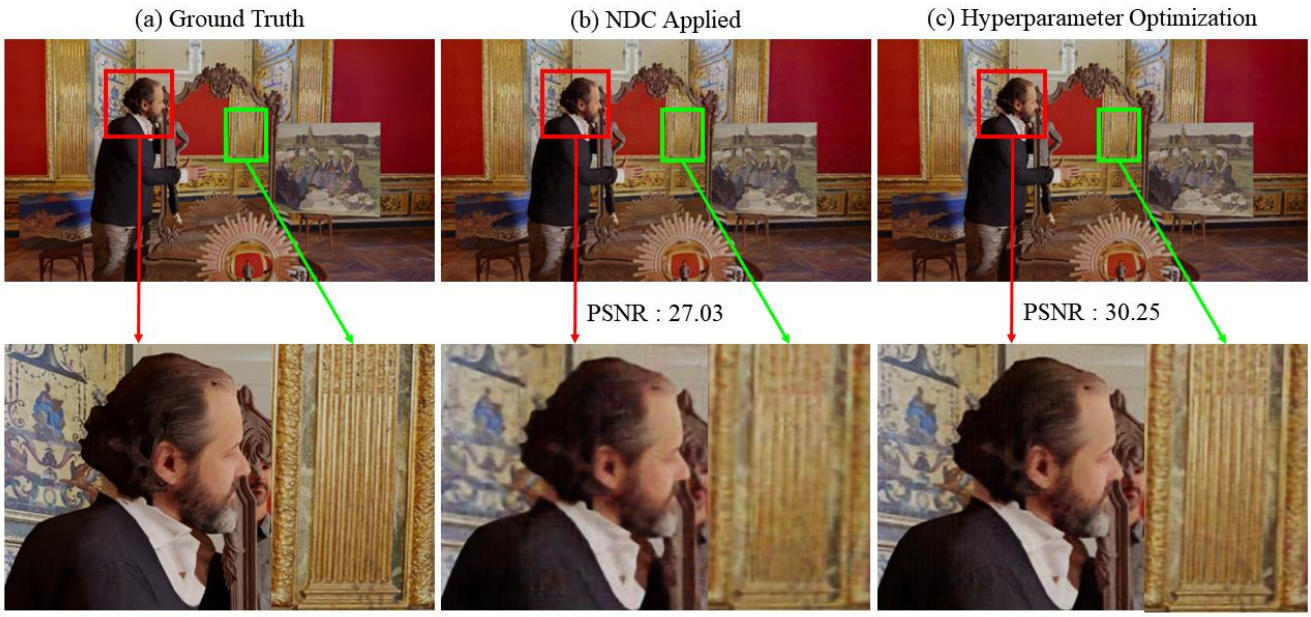

사진 2. 모델 별 렌더링 품질 결과 비교

4. 결론

본 논문에서 제안된 시스템은 기존 NeRF 기반 모델들의 한계를 극복하고 멀티뷰 동영상 및 카메라 메타데이터를 활용하여 6DoF 몰입형 비디오를 효율적으로 생성하는 새로운 프레임워크를 제시한다. 본 연구는 ReRF 모델을 확장하고, LLFF의 NDC 변환을 적용함으로써 멀티뷰 렌더링 시 깊이 정보의 왜곡을 최소화하여, 다양한 시점에서 시각적 일관성을 높였다. 실험을 통해, 본 시스템이 기존 ReRF에 비해 실사 장면 데이터 복원의 품질을 상당히 향상시킬 수 있음을 확인했다. 추가적인 하이퍼파라미터 최적화를 통해 모델의 성능을 더욱 개선시켰으며, 특히 장면 데이터의 고주파 세부 사항을 섬세하게 복원하는데 성공했다. 이러한 성과는 ReRF 모델을 복잡한 실사 장면에 적용할 수 있는 범위로 확장하고, 높은 품질의 복원을 달성함으로써 본 시스템의 유효성을 입증한다. 따라서, 본 시스템은 6DoF 비디오 생성과 같은 차세대 몰입형 컨텐츠 개발에 기여할 뿐만 아니라, 추가 연구를 통해 실시간 렌더링과 모델 압축을 적용하여 가상 현실 응용 분야에 있어 새로운 가능성을 탐색할 것으로 기대된다.

ACKNOWLEDGEMENT

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2023-00254129, 메타버스 융합대학원(성균관대학교))

참고 문헌

[1] G. Bang, Y. Liao, P. Hellier, M. Teratani, Exploration experiments on implicit neural visual representation, ISO/IEC JTC 1/SC 29/WG 04 N 0389, Geneva, July 2023.

[2] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, Vol. 65, No. 1, pp. 99-106. 2021.

[3] Sun, Cheng, M. Sun, H. T. Chen, Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, CVPR 2022.

[4] L. Wang, Q. Hu, Q. He, Z. Wang, J. Yu, T. Tuytelaars, L. Xu, M. Wu, Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos, CVPR 2023.

[5] Mildenhall, Ben, et al. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics, 38.4: 1-14. 2019.