특집논문 (Special Paper)
방송공학회논문지 제30권 제6호, 2025년 11월 (JBE Vol.30, No.6, November 2025)
https://doi.org/10.5909/JBE.2025.30.6.979
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

SplitStream: 전경-배경 분리 기반 3D 가우시안 스플래팅 스트리밍 기법

조 한 나[a)], 황 혜 민[a)], 이 장 원[a)‡]

SplitStream: Foreground–Background Separation-based 3D Gaussian Splatting Streaming Technique

Hannah Cho[a)], Hyemin Hwang[a)], and Jangwon Lee[a)‡]

요 약

본 연구는 3D Gaussian Splatting(3DGS) 기반 실시간 자유시점 비디오(FVV) 재구성에서 빈번히 발생하는 모션 블러와 고스팅 아티팩트를 줄이기 위해, Neural Transformation Cache(NTC)를 전경-배경 이중 경로로 학습하는 SplitStream을 제안한다. 첫 프레임에서 SAM2 기반 2D 마스크와 Space Carving으로 전경/배경을 분리하고, 동적 마스크를 이용한 선택적 NTC 학습 기법과 error-aware densification으로 동적 영역의 추적 정확도와 정적 영역의 시간적 안정성을 동시에 확보한다. N3DV와 MobileStage 데이터셋에서 실험한 결과, 기존 방법 대비 PSNR이 평균 +0.72dB 향상되었고, 급격한 모션 상황에서도 아티팩트 없이 선명한 객체 경계를 유지하며 안정적이고 신뢰할 수 있는 실시간 스트리밍이 가능함을 확인했다.

Abstract

We introduce SplitStream, a dual-path Neural Transformation Cache (NTC) training framework designed to reduce motion blur and ghosting artifacts in real-time Free-viewpoint Video (FVV) reconstruction with 3D Gaussian Splatting (3DGS). In the first frame, foreground and background are distinguished using SAM2-based 2D masking and space carving, after which each region is trained independently. To better capture dynamics, our method incorporates mask-guided NTC training along with error-aware densification, allowing accurate motion tracking in dynamic areas while ensuring stability in static scenes. Evaluations on the N3DV and MobileStage datasets show that SplitStream improves PSNR by +0.72dB on average over streaming baselines; even under fast motion, it preserves clear object boundaries and enables consistent, artifact-free real-time streaming.

Keyword : 3D Gaussian Splatting, Free-viewpoint Video (FVV), Foreground-Background Separation, Neural Transformation Cache (NTC), Real-time Streaming

Ⅰ. 서 론

디지털 공간에서의 참여적 상호작용이 증가함에 따라, 자유시점 비디오(Free-viewpoint Video, FVV) 재구성은 임의의 관점에서 3D 장면을 탐색할 수 있게 하는 핵심 기술로서 주목받고 있다. 이러한 3D 재구성 기술은 메타버스, 몰입형 미디어, 스포츠 중계 등 다양한 응용 분야로 확장되고 있다. 그러나 실시간으로 동적 장면을 재구성하기 위해서는 높은 시각적 품질과 빠른 처리 속도를 만족해야 하므로, 이는 여전히 해결하기 어려운 과제로 남아있다.

Neural Radiance Fields(NeRF)[1]는 암시적 표현을 통해 정적 장면에 대한 높은 재구성 품질을 달성했지만 높은 계산 부하와 느린 처리 속도로 인해 실시간 활용이 제한된다. 3D Gaussian Splatting(3DGS)[2]은 명시적 3D 가우시안 표현을 통해 이미지 평면에서 직접 splatting을 수행하여, NeRF의 한계점을 보완하며 짧은 시간 내에 고품질 이미지를 렌더링할 수 있게 되었다.

이러한 기술은 이후 동적 장면으로의 확장 연구가 활발히 진행되었다. Dynamic 3DGS[3]와 4DGS[4]는 동적 비디오를 표현하기 위해 시간 축을 추가했으나, 전체 시퀀스를 사전 학습해야 하므로 계산량이 크게 증가하는 문제가 공존했다. 한편, 3DGStream[5]은 Neural Transformation Cache(NTC)를 사용하여 메모리 부담 없이 긴 비디오를 학습하기 위해 프레임 단위의 온라인 학습 구조를 사용했다. NTC는 이전 프레임에서 학습된 가우시안의 변환 정보를 캐시 형태로 저장하고, 이를 다음 프레임 학습 시 재활용함으로써 프레임 간 일관성을 유지하면서도 전체 시퀀스를 메모리에 상주시킬 필요가 없는 구조이다. 그러나 모든 가우시안에 동일한 변환이 적용되기 때문에, 빠르게 이동하는 전경 객체가 존재할 경우 배경 영역까지 영향을 받아 모션 블러와 같은 시각적 아티팩트가 발생하는 한계가 있었다.

따라서, 본 연구는 전경과 배경을 분리해 독립적으로 학습하는 이중 경로 학습 프레임워크 SplitStream을 제안한다. SplitStream은 SAM2 기반 2D 마스크[6]와 Space Carving[7] 기법을 활용하여 초기 단계에서 전경-배경 분할을 수행한 뒤, 각 영역을 개별적으로 최적화한다. 이를 통해 전경은 동적 변화에 빠르게 적응할 수 있고 배경은 시간적 안정성을 유지할 수 있다. 이후 두 스트림을 병합하여 추가적인 움직임이나 새로 나타난 객체에 대해 보정하는 과정을 거쳐 장면의 전체적인 일관성을 높인다. 본 연구의 주요 기여는 다음과 같다. (1) 전경과 배경을 분리해 각각 최적화할 수 있는 이중 경로 NTC 기반 스트리밍 프레임워크를 제안하고, (2) 동적 마스크 기반 선택적 NTC 학습과 error-aware densification의 결합을 통해 불필요한 변형과 점 증가를 억제하며, (3) 평균 motion intensity가 높은 장면일수록 성능 향상 폭이 커짐을 정량적으로 입증하였다.

마지막으로, 본 연구에서는 동적 장면 재구성에서 주로 사용되는 N3DV[8] 데이터셋과 큰 모션을 가지고 있는 MobileStage[9] 데이터셋을 대상으로 SplitStream을 평가하였다. 실험 결과 SplitStream은 빠른 움직임이 포함된 장면에서도 PSNR 기준 평균 +0.72dB 향상을 보였으며, 이는 전경-배경 독립적 최적화의 영향으로 분석된다.

Ⅱ. 관련 연구

Free-viewpoint Video 재구성은 임의의 시점에서 동적인 장면을 사실적으로 합성하는 기술로, 대표적인 접근 방식으로는 NeRF 기반과 3DGS 기반이 있다. 각각은 고유의 장점을 지니고 있으며, 최근에는 3DGS 기반의 연구가 더 활발히 진행되고 있다.

1. NeRF 기반 방법

NeRF[1]는 3차원 좌표와 관찰 방향을 색상과 밀도로 매

a) 성균관대학교 실감미디어공학과(Department of Immersive Media Engineering, Sungkyunkwan University)
‡ Corresponding Author : 이장원(Jangwon Lee)
E-mail: leejang@skku.edu
Tel: +82-2-760-0557
ORCID: https://orcid.org/0000-0002-6601-7302
※ 이 논문의 결과 중 일부는 한국방송·미디어공학회 2025년 하계학술대회에서 발표한 바 있음
※ 이 논문은 2025년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-01-018-05)
· Manuscript September 16, 2025; Revised October 29, 2025; Accepted October 30, 2025.

핑하는 암묵적인 표현을 통해 장면을 모델링하는 방식이다. 이러한 접근 방식은 정적 장면에서 충실도도 높지만, 느린 렌더링 속도와 높은 계산 비용으로 인해 실시간 활용을 하기에는 제약이 따른다. 이러한 한계를 극복하기 위해 다양한 추가 연구가 제안되어 왔다. DyNeRF[8]는 변형 필드를 활용하여 동적 장면을 모델링하였으며, HyperReel[10]은 광선 조건부 샘플링과 키프레임 기반 볼륨 최적화를 통해 시점 의존적 렌더링 품질을 향상시켰다. 실시간 NeRF 계열 연구에는 프레임 단위 온라인 학습을 지원하는 StreamRF[11], 잔차 필드 예측을 통해 수렴 속도를 가속화하는 ReRF[12] 등이 있다. 한편 2023년에는 3DGS[2]가 제안되어 NeRF 기반 접근법의 본질적 한계를 극복하였다.

2. 3DGS 기반 방법

3DGS[2]는 3차원 가우시안을 활용한 명시적 포인트 기반 표현을 통해 정적 장면에서 빠르고 고품질의 렌더링을 가능하게 한다. 그러나 초기 기법은 시공간 동역학을 고려하지 않아 실제 변화하는 장면에 적용하는 데 한계가 있다. 이를 해결하기 위해 동적 장면을 대상으로 한 다양한 확장 연구가 진행되었다. Dynamic 3DGS[3]는 시간에 따라 변화하는 가우시안을 도입해 움직임을 처리하지만, 프레임 간 시간적 일관성이 부족하고 메모리 사용량이 크다는 단점이 있다. 4DGS[4]는 4차원 디코더를 결합하여 시공간적 일관성을 보다 효과적으로 모델링하였으며, 4K4D[9]는 고밀도 4D 그리드 특징을 활용해 4K 해상도에서 200 FPS 이상의 초고속 렌더링을 달성했지만 높은 메모리 소모가 뒤따른다. Temporal Gaussian Hierarchy[13]는 가우시안을 계층적으로 구성하여 긴 시퀀스를 일정한 GPU 메모리 사용량으로 렌더링할 수 있도록 하였다.

실시간 스트리밍 문제를 다룬 연구도 다수 존재한다. 3DGStream[5]은 NTC를 활용하여 프레임 단위로 최적화된 온라인 학습 구조를 도입하여 기존보다 더 효율적인 업데이트 방식을 지원하였다. HiCoM[14]은 가우시안 간 계층적 모션 공유를 통해 시간적 안정성을 향상시켰다. 본 연구는 3DGStream 파이프라인을 확장하여 전경과 배경을 분리 학습하는 이중 경로 전략을 제안한다. 이를 통해 배경은 시간적 안정성을 확보하고, 전경은 빠른 움직임에 정밀하게 대응한다.

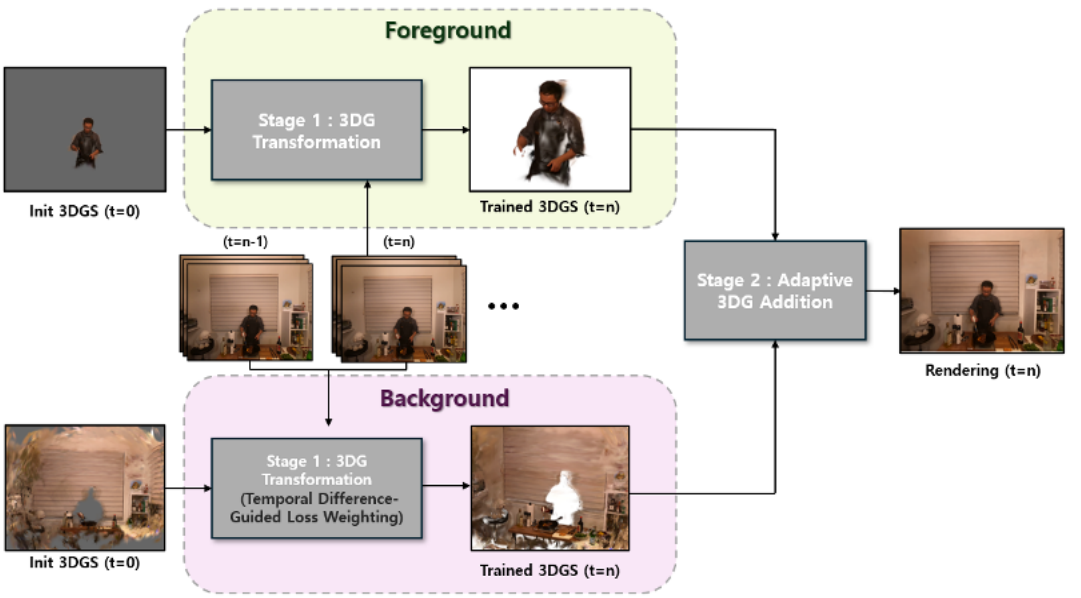


그림 1. SplitStream 3D 가우시안 스플래팅 파이프라인
Fig. 1. Pipeline of SplitStream 3D Gaussian Splatting

Ⅲ. 제안 방법

1. 개요

앞장에 언급되었듯이, 실시간 자유시점 장면에서는 빠르게 움직이는 객체로 인해 고스팅이나 번짐과 같은 아티팩트가 쉽게 발생한다. 이를 해결하기 위해 본 연구는 3DGStream[5]을 이중 스트림 구조로 확장한 SplitStream을 제안한다. 전경과 배경의 가우시안을 분리하여 독립적으로 최적화함으로써, 동적 영역과 정적 영역의 학습 목표를 분명히 한다.

그림 1은 전체 파이프라인을 보여주고 있다. 초기화는 첫 프레임에서 다중 시점 마스크를 이용해 각각 전경과 배경 3DGS 모델을 각각 생성하는 것으로 시작한다. 이후 두 모델은 3DGStream과 동일한 2단계 파이프라인을 따르되, 영역의 특성에 맞춰 기법을 조정한다. Stage 1에서는 전경과 배경을 각각 학습하며, 동적 마스크 기반 선택적 NTC 학습으로 변화가 필요한 위치에만 변환을 적용해 불필요한 가우시안 이동을 억제한다. Stage 2에서는 분리 학습된 가우시안을 병합한 뒤, Error-aware Densification으로 오차가 큰 지점에만 가우시안을 추가해 불필요한 점 증가를 방지한다.

2. 3DGStream

3DGStream은 프레임 단위의 학습을 통해 실시간 자유시점 비디오를 재구성한다. 초기 프레임에서는 COLMAP을 활용하여 희소 포인트 클라우드로부터 3D 가우시안 집합을 초기화하며, 이 집합은 이후 프레임으로 전파된다. 이후 각 프레임에서 Stage 1은 NTC(Neural Transformation Cache)를 이용해 가우시안의 변형을 추정하여 프레임 간의 변화를 포착한다. NTC는 이전 프레임의 가우시안 이동 및 회전 정보를 임시로 저장한 뒤, 이를 다음 프레임에서 초기 추정값으로 재활용하는 방식이다. 이 구조 덕분에 네트워크는 모든 프레임을 동시에 학습하지 않고도 연속적인 시간 흐름을 안정적으로 모델링할 수 있다. Stage 2에서는 view-space gradient가 큰 영역에 새로운 가우시안을 추가하여 이전에 관측되지 않은 콘텐츠를 포착한다. 단, 이렇게 추가된 가우시안은 현재 프레임에만 사용되고 다음 프레임으로는 전파되지 않는다. 그러나 이러한 접근은 빠른 움직임이 발생하는 장면에서 품질이 저하된다는 한계점이 있다. 이를 직접 실험한 결과, 그림 2에서 볼 수 있듯이, Optical flow로 측정한 카메라 움직임이 클수록 PSNR이 급격히 감소한다. 이러한 문제의식에서 출발해, 본 연구는 전경과 배경의 최적화를 분리함으로써 높은 움직임 환경에서도 일관된 재구성 성능을 유지하고자 한다.

3. 전경-배경 초기 3DGS 생성

초기 전경 및 배경 3DGS 모델을 구성하기 위해, 먼저 초기 프레임에서 사용 가능한 모든 시점의 입력 이미지를 대상으로 고해상도 Semantic Segmentation을 수행한다. 구체적으로, 각 뷰에서 전경 마스크를 생성하기 위해 SAM2[6]을 사용하였다. 다음 단계에서는 각 시점에서 얻은 2D 전경 마스크를 3차원 공간으로 역투영한다. 이 과정에서는 알려진 카메라 내·외부 파라미터를 활용하여 2D에서 전경으

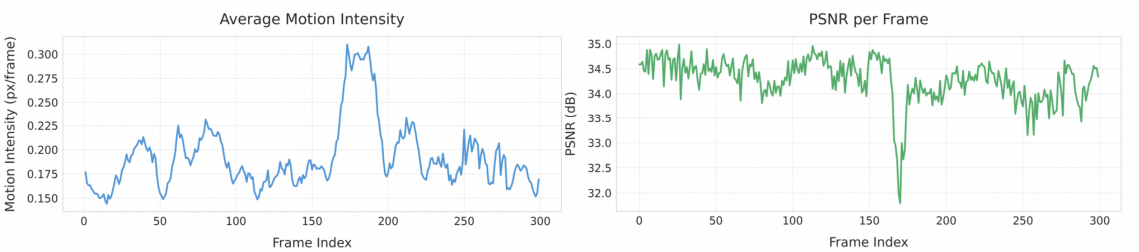


그림 2. N3DV Flame Steak 장면에서의 3DGStream 기준 기법 성능 분석 결과
Fig. 2. Performance analysis of the baseline method (3DGStream) on the N3DV Flame Steak scene

로 분류된 픽셀을 대응되는 3D 볼륨 상에 매핑한다. 모든 뷰의 정보를 집계하여 3D Voxel Grid를 구성하고, 각 voxel이 전경과 일관되는지를 평가한다.

전경과 배경을 3차원 공간에서 안정적으로 분리하기 위해, 다중 뷰 일관성을 활용한 Space Carving[7] 기법을 적용하였다. Space Carving은 여러 시점의 마스크 정보를 종합해, 모든 뷰에서 불일치하는 voxel을 제거함으로써 일관된 3D 형상을 복원하는 방법이다. 이를 통해 전경과 배경이 뷰 일관성을 유지한 형태로 분리된 3D 공간을 얻을 수 있으며, 그림 3은 Space Carving을 통해 전경과 배경이 분리된 결과를 보여준다.

이후 이 분리된 3D 공간을 활용하여 두 개의 독립적인 3DGS 모델을 초기화하였다. 전경 3DGS는 Space Carving으로 확인된 동적 영역 내부 voxel을 이용해 구성하고, 배경 3DGS는 그 외부 영역의 voxel로부터 생성하였다. 이러한 초기화는 첫 프레임에서 한 번만 수행되며, 이후 프레임에서는 두 모델을 기반으로 각각 독립적인 학습이 진행된다. Stage 1에서는 전경과 배경 모델이 각각 최적화되고, Stage 2에서는 두 모델을 병합하여 단일 가우시안 집합을 형성한다. 병합된 모델은 공동 최적화를 통해 Stage 1에서 포착되지 않은 잔여 콘텐츠나 새롭게 등장한 객체를 보완할 수 있다.

4. 이중 경로 최적화를 위한 학습 전략

4.1 Dynamic Mask 기반 선택적 NTC 학습

Stage 1에서는 전경과 배경 가우시안을 독립적으로 학습하여 각 영역의 특성에 맞는 최적화를 수행한다. 전경은 작은 범위 안에서도 객체들의 움직임이 크지만, 배경은 장면의 대부분을 차지하면서 국소적인 변화만 포함한다. 그러나 배경에서는 대부분의 영역이 정적임에도 불구하고 NTC가 전체 영역에 동일한 변형을 적용하려는 경향이 있어, 실제 움직임이 없는 부분에서도 불필요한 가우시안 이동이 발생하는 문제가 있었다.

이를 해결하기 위해 본 연구에서는 Dynamic Mask를 도입하였다. 이전 프레임 I_{t-1}과 현재 프레임 I_t의 RGB 차이를 계산하여, 변화량이 임계값 τ를 초과하는 영역만 동적으로 분류한다. 즉, 픽셀 p에 대한 마스크 $M_t(p)$는 다음과 같이 정의된다.

$$M_t(p) = \begin{cases} 1, & \text{if} \|I_t(p) - I_{t-1}(p)\| > \tau \\ 0, & otherwise \end{cases} \quad (1)$$

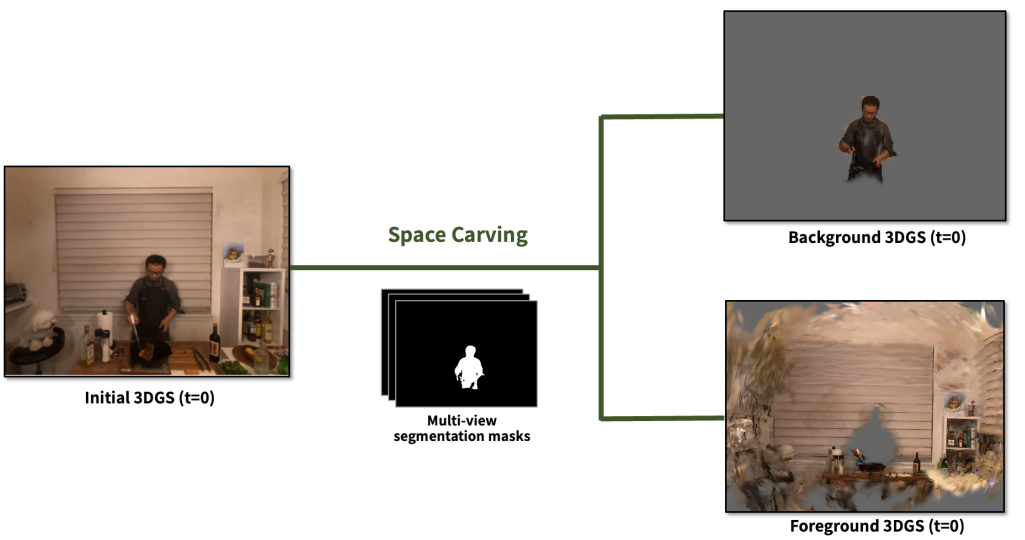


그림 3. Space Carving을 통한 전경·배경 분리 결과
Fig. 3. Foreground-Background Separation Results via Space Carving

이때 $M_t(p) = 1$인 영역만 NTC 학습이 적용되고, 정적인 영역은 이전 프레임의 위치와 속성을 그대로 유지한다. 이를 통해 정적인 영역에서 발생하는 불필요한 가우시안 이동을 효과적으로 억제하여 품질을 안정적으로 유지할 수 있다.

4.2 Error-aware Densification

Stage 2는 Stage 1에서 학습된 전경과 배경 모델을 통합한 뒤, 잔여 오차를 보정하는 과정으로 구성된다. 전경과 배경 가우시안을 병합하여 단일 모델을 형성한 후, 오차가 큰 영역에만 새로운 가우시안을 추가한다. 기존 방식은 단순히 view-space gradient를 기준으로 가우시안을 추가하여, 실제로는 변화가 없는 영역에서도 불필요한 점이 생성되는 문제가 있었다. 본 연구에서는 렌더링된 이미지 $\hat{I}_t$와 입력 이미지 I_t 간의 차이를 계산한 Error Map을 기반으로 선택적 가우시안 추가를 수행한다.

$$E_t(p) = \|I_t(p) - \hat{I}_t(p)\| \quad (2)$$

오차 $E_t(p)$가 임계값 η보다 큰 영역만을 대상으로 새로운 가우시안을 추가하여, 전경-배경 경계의 잔여 오차를 보완하고 불필요한 연산을 줄인다. 본 기법의 화질 향상은 단순한 가우시안 증가가 아니라, 오차 기반 선택적 densification을 통한 효율적인 학습 자원 배분의 결과로 나타난다.

4.3. Loss Function

본 연구는 3DGStream에서 사용된 손실 함수를 기반으로 학습을 진행하였다. Stage 1과 Stage 2 모두 렌더링된 이미지의 실제 입력 이미지 간의 화질 차이를 최소화하는 재구성 손실을 중심으로 최적화되며, 3DGStream과 동일하게 L1 + (1-SSIM) 복합 손실을 사용하며, FG/BG 각각 독립 계산 후 평균하여 최적화한다. 손실 함수는 다음과 같이 정의된다.

$$L = (1-\lambda)L_1 + \lambda L_{D-SSIM}, \; where\; \lambda = 0.2 \quad (3)$$

여기서 L_1은 렌더링된 이미지와 실제 이미지 간의 절대 오차이며, L_{D-SSIM}은 구조적 유사도(Structural Similarity, SSIM)의 손실 형태를 의미한다. SplitStream에서는 이 손실 구조를 변경하지 않고, 전경과 배경 모델 각각에 동일한 손실을 독립적으로 계산한 후 평균하여 최적화하였다. 이를 통해 기존 3DGStream의 학습 체계를 유지하면서도 전경, 배경 특성 차이를 반영할 수 있었다.

Ⅳ. 실험

1. 데이터셋

본 연구에서는 제안 기법의 일반화 성능을 확인하기 위해 서로 다른 특성을 가진 두 데이터셋을 사용하였다. 먼저, N3DV[8]는 6개 다중 뷰 시퀀스로 구성되어 있으며, 각 장면은 21개의 카메라를 통해 2704×2028 해상도로 촬영되었다. 이 중 빠른 불꽃 움직임과 반사면이 있는 Flame Steak 시퀀스를 비교에 활용하였다.

또한, 보다 큰 움직임과 인체 동작을 포함한 환경을 다루기 위해 MobileStage[9] 데이터셋의 Dance 시퀀스를 추가로 실험에 사용하였다. 해당 영상은 24개의 카메라로 1920×1080 해상도로 촬영되었다.

데이터셋 간 움직임 차이를 정량적으로 제시하기 위해 Optical Flow 기반 Motion Intensity(프레임 간 흐름 벡터 크기의 평균, 해상도 및 길이 정규화)를 산출하였다. 그 결과, Dance 시퀀스는 Flame Steak 대비 1.66배 높은 Motion Intensity를 보여, 보다 높은 난이도의 전경 모션을 포함하고 있음을 확인하였다. 이러한 정적, 동적 특성을 가진 두 시퀀스에 대한 자세한 정보는 표 1에 정리하였다.

표 1. 본 연구에서 사용한 데이터셋의 특성 비교
Table 1. Comparison of dataset characteristics

Dataset	Resolution	# of Frames	# of Cameras	Motion Intensity
N3DV[8]	2704×2028	300	21	0.19
MobileStage[9]	1920×1080	300	24	0.32

2. 구현 세부 사항

시스템은 PyTorch로 구현되었으며, NVIDIA RTX 3060

Ti GPU에서 최적화되었다. 첫 프레임의 경우 표준 3DGS 방식으로 15,000회의 반복 학습을 통해 초기 가우시안 모델을 생성한다. 이후 프레임에서는 각 프레임당 Stage 1, 150회, Stage 2, 100회, 총 250회의 최적화 과정을 수행한다. 최종 성능 지표로는 300 프레임에 대한 PSNR 평균을 사용하였다.

3. 실험 결과

표 2와 표 3에 따르면, SplitStream은 비교 대상 기법 중 가장 높은 성능을 달성하였다. 정성적 평가에서는 데이터셋별로 다른 특성을 확인할 수 있는데, 그림 4의 N3DV 데이터셋의 경우 움직임이 제한적이어서 시각적 차이가 크지 않았다. 반면 그림 5의 MobileStage Dance 시퀀스의 경우, 큰 움직임이 포함되어 있어 제안 기법의 효과를 뚜렷하게

표 2. N3DV Flame Steak 장면에서의 정량적 결과 (*표시는 우리와 동일한 실험 환경에서 공식 코드를 사용하여 수행된 평가를 의미함)
Table 2. Quantitative Results on the N3DV Flame Steak scene (*indicates that the evaluation was performed using the official code in the same experimental environment as ours)

Method	PSNR (dB↑)	Train (s↓)
StreamRF[11]	32.09	15
HiCoM[14]	31.17	6.7
3DGStream*[5]	33.11	34
SplitStream*	33.83	39.4

표 3. MobileStage Dance 장면에서의 정량적 결과 (*표시는 우리와 동일한 실험 환경에서 공식 코드를 사용하여 수행된 평가를 의미함)
Table 3. Quantitative Results on MobileStage Dance scene (*indicates that the evaluation was performed using the official code in the same experimental environment as ours)

Method	PSNR (dB↑)	Train (s↓)
3DGStream*[5]	26.93	58.49
SplitStream*	27.38	69.54

볼 수 있었다. 3DGStream의 경우 모션 블러와 아티팩트가 발생한 반면, SplitStream의 경우 불필요한 가우시안의 추가를 방지하여 아티팩트가 줄어들어 더 선명한 결과를 얻을 수 있었다. 이러한 결과는 제안한 방법이 큰 움직임이 있는 장면에서 특히 효과적임을 입증한다. 실시간 추론(렌더링)은 유지되며, 프레임당 업데이트 지연은 3DGStream 대비 +16% 수준이다. 이는 FG/BG 순차 최적화로 인한 오버헤드로, 병렬화 시 해소 가능하다.

또한, 표 4에서 제안 기법이 전경-배경에 각각 어떤 영향을 미치는지 분석하기 위해, 3DGStream과 SplitStream을 학습한 후 동일 전경 및 배경 마스크로 PSNR을 각각 평가하였다. 실험 결과, 대규모 전경 움직임을 포함하는 MobileStage에서는 전경 PSNR이 기존 대비 +1.28dB 향상되었다. 이는 기존 통합 최적화 방식에서 전경의 위치 변화가 배경으로 누적되어 발생하는 고스팅 및 모션 블러를 효과적으로 억제하였음을 의미한다. 반면 N3DV는 움직임이 비교적 제한적이기 때문에 개선 폭이 상대적으로 작게

그림 4. N3DV Flame Steak 장면에서의 정성적 비교 결과
Fig. 4. Qualitative Results on N3DV Flame Steak scene

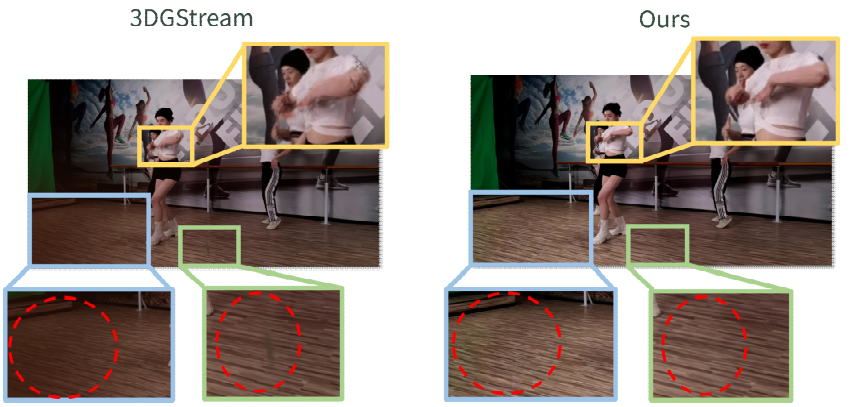


그림 5. MobileStage Dance 장면에서의 정성적 비교 결과
Fig. 5. Qualitative Results on MobileStage Dance scene

표 4. Ablation: N3DV, MobileStage에서의 전경-배경 분리 성능
Table 4. Ablation: Separate Foreground(FG)-Background(BG) Performance on N3DV, MobileStage scenes

Dataset	Method	FG PSNR (dB ↑)	BG PSNR (dB ↑)
N3DV[8]	3DGStream[5]	31.19	33.29
	SplitStream	32.34 (+0.15)	33.93 (+0.64)
MobileStage[9]	3DGStream[5]	22.15	28.37
	SplitStream	23.43 (+1.28)	28.46 (+0.09)

나타났다. 특히 MobileStage의 평균 Motion Intensity가 N3DV 대비 약 1.66배 높다는 점(표 1 참조)을 감안하면, 이러한 결과는 제안한 분리 최적화가 Motion Intensity가 높은 장면에서 특히 효과적임을 보여준다.

한편, 두 데이터셋 모두에서 배경 PSNR이 전경보다 높게 나타난 것은 배경 영역이 정적인 변화 위주로 구성되어 오차 누적 가능성이 적기 때문이다. 요약하면, 분리 학습 구조는 전경과 배경이 상호 간섭 없이 자체적 수렴을 달성하도록 돕는 구조적 안정성을 제공하며, 이는 동적인 환경에서 더욱 두드러지는 장점이다.

V. 결 론

본 연구에서는 실시간 3D Gaussian 스트리밍 시스템에서 발생하는 모션 관련 아티팩트를 해결하기 위해 전경-배경 분리 기반의 이중 스트림 프레임워크 SplitStream을 제안하였다. SplitStream은 2D 마스킹과 공간 카빙을 통해 전경과 배경을 분리하고 이중 경로 NTC 학습을 진행하였다. 또한 학습 중 생기는 아티팩트를 동적 마스크 기반 NTC 학습과 error-aware densification을 통해 방지하였다. 실험 결과, SplitStream은 N3DV 데이터셋에서 33.83dB의 PSNR을 달성하여 화질을 높였다. 또한 동적인 장면을 포함하는 MobileStage 데이터셋에서도 모션 블러와 아티팩트를 해결하여 기존 방식 대비 안정적인 성능을 유지할 수 있었다.

하지만 제안 기법은 몇 가지 제약사항이 존재한다. 첫째, 전경과 배경을 분리하여 순차적으로 학습하였기 때문에 기존 방식에 비해 전체 연산 시간이 증가하였다. 다만, 이는 pruning 등의 경량화 기법 및 파이프라인 병렬화를 통해 완화될 수 있다. 둘째, 전경-배경 분리의 품질은 첫프레임 재구성 품질과 segmentation 마스크의 정확도에 크게 의존한다. Segmentation이 불완전할 경우 경계 부분에서 아티팩트가 발생하여 최종 재구성의 품질에 영향을 미칠 수 있다. 향후 연구에서는 마스크 자동화, 다중 전경 객체 지원, 학습 시간 단축 등을 통해 프레임 간 시간적 연결성을 높이고, 시스템의 확장성과 효율성을 개선할 예정이다.

참 고 문 헌 (References)

[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," *Communications of the ACM*, Vol. 65, No. 1, pp. 99-106, January 2021.
doi: https://doi.org/10.1145/3503250

[2] 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, Article 139, pp. 1-14, July 2023.
doi: https://doi.org/10.1145/3592433

[3] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, "Dynamic 3D Gaussians: Tracking by persistent dynamic view synthesis," *Proceedings of International Conference on 3D Vision*, Davos, Switzerland, pp. 800-809, 2024.
doi: https://doi.org/10.1109/3DV62453.2024.00044

[4] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4D Gaussian splatting for real-time dynamic scene rendering," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 20310-20320, June 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01920

[5] J. Sun, H. Jiao, G. Li, Z. Zhang, L. Zhao, and W. Xing, "3DGStream: On-the-fly training of 3D Gaussians for efficient streaming of photo-realistic free-viewpoint videos," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 20675-20685, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01954

[6] N. Ravi, V. Gabeur, Y. T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," *arXiv preprint*, arXiv:2408.00714, 2024.
doi: https://doi.org/10.48550/arXiv.2408.00714

[7] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision*, Vol. 38, No. 3, pp. 199-218, July 2000.
doi: https://doi.org/10.1023/A:1008191222954

[8] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, and Z. Lv, "Neural 3D video synthesis from multi-view video," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 5521-5531, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00544

[9] Z. Xu, S. Peng, H. Lin, G. He, J. Sun, Y. Shen, H. Bao, and X. Zhou, "4K4D: Real-time 4D view synthesis at 4K resolution," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 20029-20040, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01893

[10] B. Attal, J. B. Huang, C. Richardt, M. Zollhoefer, J. Kopf, M. O'Toole, and C. Kim, "HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 16610-16620, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.01594

[11] L. Li, Z. Shen, Z. Wang, L. Shen, and P. Tan, "Streaming radiance fields for 3D video synthesis," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 13485-13498, December 2022.
doi: https://doi.org/10.48550/arXiv.2210.14831

[12] L. Wang, Q. Hu, Q. He, Z. Wang, J. Yu, T. Tuytelaars, L. Xu, and M. Wu, "Neural residual radiance fields for streamably free-viewpoint videos," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 76-87, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.00016

[13] Z. Xu, Y. Xu, Z. Yu, S. Peng, J. Sun, H. Bao, and X. Zhou, "Representing long volumetric video with temporal Gaussian hierarchy," *ACM Transactions on Graphics*, Vol. 43, No. 6, Article 234, pp. 1-18, November 2024.
doi: https://doi.org/10.1145/3687919

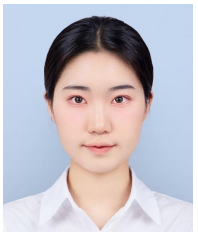
[14] Q. Gao, J. Meng, C. Wen, J. Chen, and J. Zhang, "HiCoM: Hierarchical coherent motion for streamable dynamic scene with 3D Gaussian Splatting," *Advances in Neural Information Processing Systems*, Vol.37, pp.80609-80633, Dec. 2024.
doi: https://doi.org/10.48550/arXiv.2411.07541

저 자 소 개

조 한 나

- 2024년 2월 : 인하대학교 정보통신공학과 학사
- 2025년 3월 ~ 현재 : 성균관대학교 실감미디어공학과 석사과정
- ORCID : https://orcid.org/0009-0006-5158-6955
- 주관심분야 : 3D Reconstruction, Human-Computer Interaction

저 자 소 개



황 혜 민

- 2023년 : 한국항공대학교 전자 및 항공전자공학과 학사
- 2023년 ~ 현재 : 성균관대학교 실감미디어공학과 석사과정
- ORCID : https://orcid.org/0009-0001-4003-2395
- 주관심분야 : 3D Reconstruction, Computer Vision

이 장 원

- 2006년 : 성균관대학교 정보통신공학부 학사
- 2008년 : 성균관대학교 전자전기컴퓨터공학과 석사
- 2018년 : Indiana University Intelligent and Interactive Systems 박사
- 2018년 ~ 2020년 : ObjectVideo Labs at Alarm.com Research Scientist
- 2020년 ~ 2023년 : 한국항공대학교 항공전자정보공학부 교수
- 2023년 ~ 현재 : 성균관대학교 실감미디어공학과 교수
- ORCID : https://orcid.org/0000-0002-6601-7302
- 주관심분야 : Computer Vision for Robotics, Human-Robot Interaction