졸업논문
소개

다시점 확산 모델을 위한 대응점-어텐션 정렬 기법 연구

권민경 / KAIST 김재철AI대학원

신규 시점 합성(Novel View Synthesis, NVS)은 주어진 참조 이미지로부터 관측되지 않은 시점의 이미지를 기하학적 일관성과 사실적인 외관을 유지하며 예측하는 문제이다. NeRF, 3D Gaussian Splatting과 같은 최적화 기반 기법은 장면마다 수십 장의 입력 이미지와 별도의 최적화 과정을 필요로 하는 반면, 최근에는 대규모 2D 확산 모델의 생성 사전지식을 활용하는 다시점 확산 모델(multi-view diffusion model)이 새로운 패러다임으로 부상하였다. CAT3D, MVGenMaster 등의 다시점 확산 모델은 각 시점의 토큰을 하나의 시퀀스로 연결하여 시점 간 정보를 교환하는 3D self-attention을 채택함으로써 시점 간에 기하학적으로 일관된 이미지를 생성한다. 그러나 시점 변화가 크거나 기하 구조가 복잡한 경우에는 시점 간 정렬 오류와 구조 왜곡이 빈번히 발생하며, 이러한 시점 일관성이 모델 내부의 어떤 메커니즘으로부터 비롯되는지는 명확히 규명되지 않았다.

본 연구에서는 다시점 확산 모델의 3D self-attention map을 분석하여 세 가지 핵심 관찰을 얻었다. 첫째, 모델은 학습 과정에서 기하학적 대응 관계(geometric correspondence)를 어텐션 계층에 자연스럽게 인코딩하며, 이 특성은 특정 계층에 집중되어 나타난다. 둘째, 학습이 진행될수록 해당 계층의 대응 정밀도와 PSNR이 함께 증가하여 대응 관계의 품질이 생성 품질과 직접적으로 연관됨을 확인하였다. 셋째, 그럼에도 이 대응 신호는 불완전하여 VGGT와 같은 기하 예측 모델 대비 정밀도 격차가 크고, 특히 큰 시점 회전에서 대응 정확도가 급격히 저하된다.

이러한 분석을 바탕으로 본 연구에서는 어텐션 맵을 기하학적 대응 관계로 직접 지도하는 학습 기법인 CAMEO(Correspondence-Attention Alignment for Multi-view Diffusion Models)를 제안한다. 먼저 VGGT와 같은 기하 예측 모델로부터 각 영상의 포인트맵(pointmap)을 획득하고, 3D 공간에서의 최근접 이웃 탐색으로 영상 쌍 간 픽셀 대응을 구한 뒤 이를 어텐션 토큰 해상도로 정렬한다. 각 쿼리 토큰에 대해 대응되는 키 토큰 위치를 1로 갖는 원-핫 벡터를 쌓아 기하 대응 맵을 구성하며, 순환 일관성(cycle consistency) 검사를 통해 가려지거나 신뢰할 수 없는 대응을 걸러내는 가시성 마스크를 함께 구성한다. CAMEO는 참조 시점과 목표 시점을 포함한 모든 시점 쌍 (i, j)에 대해 교차 시점 어텐션 맵과 기

하 대응 맵 간의 교차 엔트로피(cross-entropy) 손실을 가시성 마스크로 가중하여 계산한다. 최종 학습 목적 함수는 기존 노이즈 제거 손실에 CAMEO 손실을 더한 형태로 모델 구조의 변경 없이 기존 학습 파이프라인에 그대로 통합할 수 있다. 특히 기하 대응이 가장 강하게 나타나는 단일 어텐션 계층 하나만 지도하더라도 인접 계층까지 대응 정밀도가 함께 향상되어, 모델 전체가 정확한 시점 간 대응을 학습하도록 유도할 수 있다.

제안 기법의 효과를 검증하기 위해 CAT3D를 기본 모델로 하여 장면 수준의 RealEstate10K와 객체 중심의 CO3D 데이터셋에서 각각 학습하고, PSNR, SSIM, LPIPS로 신규 시점 합성 성능을 비교 분석하였다. 비교 대상으로는 특징 정렬 기법인 REPA와 Geometry Forcing을 동일 조건에서 학습하였으며, 학습 분포 밖(out-of-domain) 일반화를 확인하기 위해 장면 수준 모델을 DTU 데이터셋에서 추가로 평가하였다. 실험 결과, RealEstate10K에서 CAMEO는 학습 8만 회 시점에 PSNR 19.40에 도달한 반면, 기본 모델이 같은 수준에 이르기까지는 16만 회 이상의 학습이 필요하였다. 또한 대부분의 학습 구간에서 REPA와 Geometry Forcing을 상회하였고, 수렴 이후에도 세 데이터셋 모두에서 기본 모델 대비 우수한 성능을 유지하여 단순한 학습 가속을 넘어 최종 성능 자체를 향상시킴을 확인하였다.

본 연구에서는 다시점 확산 모델의 3D self-attention이 학습 과정에서 기하학적 대응 관계를 자발적으로 획득하며 이것이 시점 일관성과 생성 품질을 좌우하는 핵심 신호임을 밝히고, 동시에 이 신호가 큰 시점 변화에서 취약하다는 한계를 확인하였다. 이에 기하 대응으로 어텐션 맵을 직접 지도하는 CAMEO를 제안하여 수렴 속도를 약 2배 높이는 동시에 신규 시점 합성 품질을 개선하고, 시점 변화가 큰 경우에도 참조 영상의 기하 구조를 정확히 보존하도록 하였다. CAMEO는 교차 시점 어텐션을 사용하는 다양한 다시점 확산 모델에 손쉽게 통합될 수 있으며, 생성된 시점 영상으로 3D Gaussian Splatting을 최적화한 3D 복원 실험에서도 기본 모델이 실패하는 장면을 일관되게 복원하였다. 다만 참조 시점과 목표 시점 간 시각적 중첩이 거의 없는 극단적인 시점 변화에서는 대응 관계 자체를 정의하기 어렵다는 한계가 있으며, 이 경우 중간 시점을 순차적으로 생성하는 전략과의 결합이 대안이 될 수 있다. 향후 비디오 확산 모델 및 4D 복원으로의 확장, 그리고 다른 계층에 인코딩된 의미론적 대응(semantic correspondence)을 활용한 지도 방식이 후속 연구 방향으로 기대된다.



권 민 경

- 2025년 2월 : 포항공과대학교 컴퓨터공학과 학사
- 2025년 2월 ~ 현재 : KAIST AI 석사
- 2026년 5월 ~ 현재 : 네이버 AI Lab 인턴
- 주관심분야 : 생성형 모델, 3D/4D 생성, 월드 모델