KCGS2025

GENERATING WORLDS, RENDERING REALITY

KOREA COMPUTER GRAPHICS SOCIETY

한국컴퓨터그래픽스학회

2025 학술대회

학술대회 논문집

한국컴퓨터그래픽스 학회 2025 학술대회 논문집

2025년 7월 8일-11일
고성 델피노 리조트

KCGS2025

GENERATING WORLDS, RENDERING REALITY

주관기관: 한국컴퓨터그래픽스학회

KCGS 2025 환영사

고성 델피노 리조트에서 열리는 2025년 한국컴퓨터그래픽스학회 학술대회 및 여름학교에 여러분을 모시게 되어 매우 기쁩니다. 아름다운 자연과 첨단 기술이 어우러진 고성에서 참석해 주신 모든 분께 진심으로 환영의 말씀을 전합니다.

한국컴퓨터그래픽스학회는 학계와 산업계 연구자들이 모여 최신 기술과 창의적 아이디어를 나누며 함께 성장하는 자리입니다. 이번 학술대회는 컴퓨터그래픽스를 비롯해 게임, 특수효과, 가상·증강현실, 인간-컴퓨터 상호작용, 메타버스, 미디어아트 등 다양한 분야를 아우르며 협력과 소통으로 미래를 모색합니다. 올해 주제는 ‘Generating Worlds, Rendering Reality’이며, 명망 높은 초청강연, 우수 해외학술대회 발표, 창해 신진연구자 및 석사논문상 후보 발표, 산업체와 대학전시 등 풍성한 프로그램이 준비되어 있습니다.

특히 컴퓨터그래픽스 분야의 산학협력 활성화에 헌신하신 서울대학교 신영길 교수님께 공로패를 드리며 깊은 감사와 축하를 전합니다. 준비를 위해 애써주신 조직위원과 후원 기관 및 기업에 깊이 감사드리며, 이 행사가 여러분 모두에게 유익하고 뜻깊은 교류의 시간이 되길 바랍니다. 고성에서의 나흘 동안 따뜻한 추억과 함께 건강과 행복이 늘 함께하시길 기원합니다. 감사합니다.

2025년 7월 8일

사단법인 한국컴퓨터그래픽스학회장 최수미

학술대회 조직 위원회

대회장	최수미 (세종대)		
총무	강형업 (고려대)	김종현 (인하대)	
조직위원장	이윤상 (한양대)	최명걸 (가톨릭대)	
프로그램위원장	백승환 (POSTECH)	장 윤 (세종대)	
여름학교준비위원장	김준호 (국민대)	송오영 (세종대)	
조직위원	김성기 (조선대)	김성희 (동의대)	김승욱 (한국외대)
	손진희 (GIST)	유 리 (아주대)	허재필 (성균관대)
프로그램 위원	강형엽 (고려대)	계희원 (한성대)	경민호 (아주대)
	권구주 (배화여대)	권태수 (한양대)	김광욱 (한양대)
	김덕수 (코리아텍)	김민혁 (KAIST)	김선정 (한림대)
	김성기 (조선대)	김승욱 (한국외대)	김영민 (서울대)
	김종현 (인하대)	김준호 (국민대)	김진모 (한성대)
	김학구 (중앙대)	노준용 (KAIST)	문보창 (GIST)
	박상일 (세종대)	박상훈 (서강대)	박 준 (홍익대)
	박진아 (KAIST)	손봉수 (중앙대)	송재원 (㈜덱스터스튜디오)
	송오영 (세종대)	송현주 (숭실대)	오경수 (숭실대)
	유 리 (아주대)	윤성의 (KAIST)	윤승현 (동국대)
	이강훈 (광운대)	이성희 (KAIST)	이성길 (성균관대)
	이영호 (목포대)	이인권 (연세대)	이정진 (숭실대)
	이종원 (세종대)	이주호 (서강대)	이지은 (한성대)
	이창하 (중앙대)	이환용 (아주대)	이택희 (한국공학대)
	임인성 (서강대)	임현기 (경기대)	정문열 (서강대)
	정원기 (고려대)	정주립 (서울여대)	조동식 (울산대)
	조성현 (POSTECH)	최명걸 (가톨릭대)	최민규 (광운대)
	최유주 (SMIT)	최아영 (가천대)	최종인 (서울여대)
	한다성 (한동대)	한정현 (고려대)	황재인 (KIST)
	허재필 (성균관대)	홍헬렌 (서울여대)	

목 차

초청강연

특별세션

후원 기업 발표

7월 10일 13:40~14:20

창해 신진 연구자상 발표

7월 10일 14:35~15:35

석사 논문상 후보자 발표

7월 09일 14:20~16:00

여름학교

우수 국제 학술대회 논문 발표: 이미지/비디오

7월 09일 13:00~14:00

우수 국제 학술대회 논문 발표: 캐릭터 컨트롤

7월 09일 13:00~14:00

우수 국제 학술대회 논문 발표: 메쉬/VR/AR

7월 10일 13:40~14:20

논문 발표: 가상/증강현실

7월 10일 09:00~10:40

논문 발표: HCI/시각화/시스템

7월 10일 09:00~10:40

논문 발표: 애니메이션

7월 11일 09:00~10:20

논문 발표: 그래픽스응용

7월 11일 09:00~10:20

논문 발표: 렌더링/이미지/비디오

7월 11일 10:30~11:40

논문 발표: 시뮬레이션/모델링

7월 11일 10:30~11:40

포스터 발표

7월 10일 17:30~18:30

포스터 발표

7월 10일 17:30~18:30

초 청 강 연

초 청 강 연

AX 2.0시대 우리의 준비

홍진배 원장, IITP

강연 내용

AI 기술은 생성형 AI를 넘어 'AI 에이전트'와 '피지컬 AI'로 대표되는 'AX(AI 대전환) 2.0 시대'로 빠르게 진입하고 있다. 본 강연에서는 급변하는 글로벌 AI 트렌드를 심층 분석하고, AX 2.0 시대를 주도할 핵심 주권 기술인 AI 모델, AI 반도체, 차세대 네트워크, 사이버 보안, 양자, AI 융합서비스 기술을 조망한다. 글로벌 AI 경쟁에서 살아남기 위해 집중해야 할 준비 과제와 R&D의 전략적 방향성을 제시하고자 한다.

강연자 이력

2005-2008	University of Manchester 기술경영학 박사
2021	대통령 홍조근정훈장
2019-2022	과학정보기술통신부 정보보호네트워크정책관, 통신정책관
2022-2024	과학기술정보통신부 네트워크정책실장
2024-	정보통신기획평가원 원장

초 청 강 연

의료IT가 여는 스마트병원의 미래

최규옥 회장, 오스템임플란트

강연 내용

미래의 의료 소프트웨어는 병원 운영 전반을 아우르는 통합 관리 기능과 더불어, 환자 데이터를 실시간으로 수집하고 분석하며 AI를 기반으로 한 진찰, 진단, 처방까지 지원하게 됩니다. 이는 진료의 정밀성과 신속성을 높일 뿐 아니라, 환자 맞춤형 치료를 가능하게 하며, 의료진의 의사결정을 더욱 과학적이고 정교하게 만드는 핵심 도구로 자리매김할 것입니다.

무엇보다도 중요한 변화는, 의료 소프트웨어가 개별 의료진의 지식과 경험의 한계를 넘어, 고도로 직접된 의료 빅데이터와 AI 기술을 통해 수준 높은 전문가들의 임상 지식과 진단 노하우를 실시간으로 공유하고 활용할 수 있게 된다는 점입니다. 이는 모든 의료진들이 보다 정밀하고 신뢰도 높은 진료를 제공할 수 있게 함으로써, 진료 수준의 획기적인 향상을 가능하게 합니다.

치과 분야는 이러한 디지털 전환이 가장 빠르게 진행되고 있는 대표적인 사례입니다. '디지털 덴티스트리'는 3D 스캐너로 환자의 구강 데이터를 정밀하게 수집하고, CAD 소프트웨어를 통해 치료 계획을 시뮬레이션하며, CAD/CAM 기술을 활용한 '원데이 임플란트'와 보철 치료가 이미 보편화되고 있습니다. 아울러, 치과용 AI는 영상 판독, 교정 설계, 임플란트 플래닝 등 다양한 진료 영역에서 의사들의 의사결정을 지원하며, 경험이 부족한 초보 치과의사도 숙련된 전문가 수준의 치료를 제공할 수 있도록 돕고 있습니다.

결론적으로, AI 고정밀 영상 처리 기술, 3D 프린터, CAD/CAM 등 첨단 디지털 기술의 융합은 치과를 포함한 전 의료 분야의 서비스 질과 효율을 동시에 높이고 있으며, 이는 미래 병원과 치과의 패러다임을 근본적으로 바꾸는 핵심 동력으로 작용할 것입니다.

강연자 이력

1991	서울대학교 치의학과 졸업
1996	단국대학교 대학원 졸업 (치의학석사)
1997	오스템임플란트 설립
2010	고려대학교 대학원 졸업 (의학박사)
	한국중견기업연합회 이사

초 청 강 연

Octree-based 3D Representation and Learning

Peng Shuai Wang 교수, Peking University

강연 내용

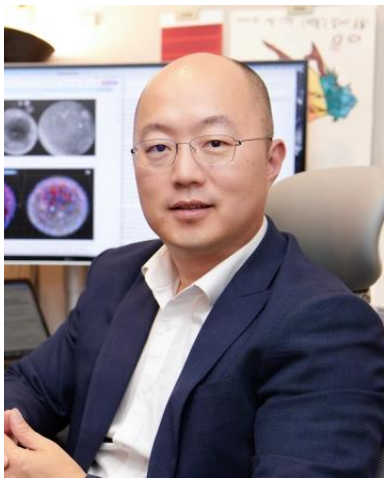
In recent years, 3D deep learning has gained significant attention in both academia and industry. However, the diversity of 3D data representations often necessitates the design of specialized neural network architectures tailored to specific shape formats and downstream tasks, which adds considerable complexity to learning systems. To address this challenge, my research focuses on developing a general and efficient framework for 3D deep learning. In this talk, I will present my recent progress toward this goal, including works on octree-based CNNs, GNNs, Transformers, and octree-based diffusion models and GPTs. We expect this unified framework to bridge the gap across different 3D representations and tasks, and to advance the development of general-purpose 3D intelligent models.

Peng-Shuai Wang is currently an Assistant Professor at Peking University. His research interests lie in computer graphics, geometry processing, and 3D deep learning. He serves as an Associate Editor for IEEE TVCG and Computers & Graphics, and as a Program Committee member for several major international graphics conferences, including SIGGRAPH Asia, Eurographics, SGP, and CVM. He received the Asiagraphics Young Researcher Award in 2023 and the China3DV Excellent Young Researcher Award in 2025.

강연자 이력

2018	Ph.D. from the Institute for Advanced Study at Tsinghua University
2018-2022	Senior researcher in the graphics group at Microsoft Research Asia
2023	AsiaGraphics Young Researcher Award 수상
2025	China3DV Excellent Young Researcher Award 수상
2022-	Assistant Professor at Peking University

초 청 강 연



Holotomography and virtual staining: interference of 3D fluorescence and H&E images from abel-free samples

박용근 교수, KAIST

강연 내용

Holotomography (HT) is a powerful label-free imaging technique that enables high- resolution, three-dimensional quantitative phase imaging (QPI) of live cells and organoids through the use of refractive index (RI) distributions as intrinsic imaging contrast 1-3 . Similar to X-ray computed tomography, HT acquires multiple two-dimensional holograms of a sample at various illumination angles, from which a 3D RI distribution of the sample is reconstructed by inversely solving the wave equation. By combining label-free and quantitative 3D imaging capabilities of HT with machine learning approaches, there is potential to provide synergistic capabilities in bioimaging and clinical diagnosis. In this presentation, we will discuss the potential benefits and challenges of combining QPI and artificial intelligence (AI) for various aspects of imaging and analysis, including segmentation, classification, and imaging inference 3-6 . We will also highlight recent advances in this field and provide insights on future research directions. Overall, the combination of QPI and AI holds great promise for advancing biomedical imaging and diagnostics.

강연자 이력

2004-2005	Research Assistant, Institute of Advanced Machinery and Design, Seoul National University
2005-2010	Research Assistant, Laser Biomedical Research Center, MIT
2010	Visiting Scholar, MIT
2010-	Associate Professor, KAIST

초 청 강 연

전략적 데이터의 시대: 생성형 AI와 합성데이터의 산업적 전환점

조호진 대표, 젠젠에이아이

강연 내용

최근 수년 간 AI 연구는 눈부신 속도로 발전해 왔으며, CNN부터 Transformer까지 다양한 모델이 발표되며 산업 전반에 걸쳐 기대를 모으고 있다. 그러나 실제 산업 현장에서 AI가 제대로 작동하지 않는 가장 큰 이유는 여전히 데이터의 품질과 적합성이다. 모델의 성능은 이제 한계에 가까워졌고, AI의 다음 진화는 데이터를 어떻게 확보하고, 설계하고, 적용하느냐에 달려 있다. 이러한 흐름 속에서 합성데이터(synthetic data)는 데이터 수집의 한계를 극복하고, 특정 목적에 최적화된 학습 환경을 구축하는 핵심 도구로 떠오르고 있다. 특히, 생성형 AI 기반의 데이터 생성 기술은 기존 방식으로는 수집이 어렵거나 위험한 상황(예: 드문 사고 장면, 민감한 의료 이미지, 고객 개인정보)을 대체하고 보완할 수 있는 새로운 가능성을 열어주고 있다.

본 강연에서는 생성형 AI의 발전과 함께 변화하고 있는 데이터 전략의 패러다임을 조망하고, GenGenAI가 실제 산업 현장에서 수행한 합성데이터 구축 사례를 중심으로 그 실효성과 성능 향상 효과를 공유한다. 자율주행, 국방, 보안 등 주요 고객 사례를 통해 합성데이터가 어떻게 실제 AI 시스템의 성능과 신뢰성을 향상시키고 있는지를 살펴보고, 데이터 중심 AI 시대에 있어 산업과 학계가 함께 고민해야 할 데이터 다양성 확보에 대한 비전을 제안한다.

강연자 이력

포항공과대학교 컴퓨터 공학과 학사 졸업

포항공과대학교 컴퓨터 공학과 박사 졸업

Adobe Photoshop 개발팀 인턴 근무

스타트업 스트라드비젼 근무

젠젠에이아이 창업

창해 신진 연구자상 발표

2025 창해 신진 연구자상 후보

박정남 (Jungnam Park)

jungnam04@imo.snu.ac.kr

박정남(Jungnam Park)은 2017년 2월 POSTECH에서 컴퓨터공학 학사학위를, 2025년 2월 서울대학교에서 컴퓨터공학 박사학위를 취득했으며, 현재 서울대학교 지능형 동작 연구실에서 박사후연구원으로 재직 중이다. 그의 주요 연구 분야는 컴퓨터 애니메이션으로, 특히 물리 기반 근골격 시뮬레이션과 심층강화학습을 활용해 사실적인 인간 동작을 생성하고 분석하는 데 초점을 맞추고 있다.

그는 물리 기반 시뮬레이션 상에서, 실제 사람과 유사한 뼈와 근육 구조를 갖춘 근골격 모델을 제어함으로써, 해부학적 특성과 움직임 사이의 상관관계를 규명하는 연구를 수행해왔다. 수백 개의 근육 및 관절 상태에 따른 보행 동작 변화를 예측하는 심층강화학습 기반의 제어기를 제안하고, 반대로 사람의 보행 동작으로부터 근육과 뼈의 상태를 추정하는 연구도 진행하였다. 또한, 대량의 동작 데이터를 모방할 수 있는 근육 제어기를 학습하여, 사람의 동작이 주어졌을 때 각 근육이 얼마나 사용되는지(근활성도)를 예측할 수 있는 모델도 개발하였다.

이러한 연구 성과를 바탕으로 그는 SIGGRAPH 2022, 2023에 각각 한 편의 논문 발표 및 SIGGRAPH 2025에 한 편의 논문이 채택되어 캐릭터 애니메이션 분야에 학술적 기여를 하였다. 이 외에도 그는 SIGGRAPH 및 SIGGRAPH Asia에 총 네 편의 논문에 공저자로 참여하여 관련 분야의 지속적인 연구 발전에도 힘을 보탰다. 또한 그는 이러한 연구 경험을 바탕으로 NeurIPS MyoChallenge 2023 Locomotion Track에서 1위를 차지하였으며, 이를 통해 biomechanics 및 AI 분야의 연구 커뮤니티에 성과를 공유하고 기여하였다. 2024년에는 Meta Reality Labs에서 인턴십을 수행하며, 산업 분야에 적용 가능한 기술 개발에도 참여하였다.

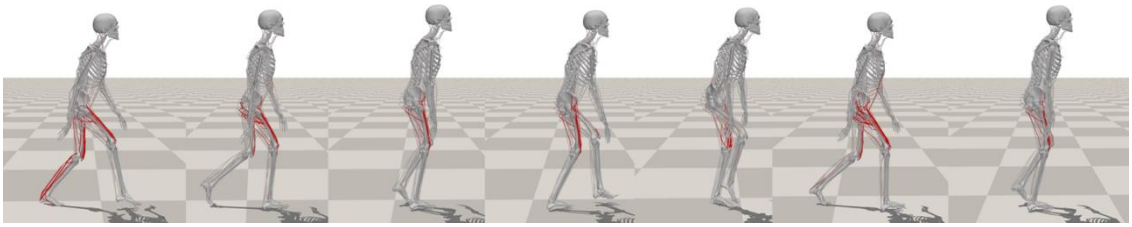


그림 1. 근골격 시뮬레이션을 통해 생성한 병증을 가진 보행

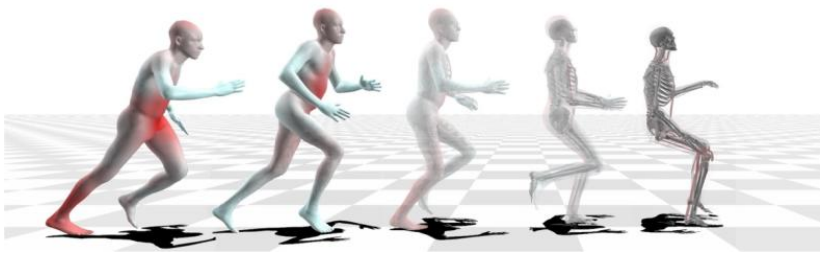


그림 2. 동작을 모방하는 근골격 모델을 통한 근활성도 예측

2025 창해 신진 연구자상 후보

배진석 (Jinseok Bae)

capoo95@snu.ac.kr

배진석 (Jinseok Bae)은 서울대학교 전기·정보공학부 박사과정생으로, 컴퓨터 그래픽스 및 비전 분야 중에서도 캐릭터 애니메이션을 중심으로 연구하고 있다. 연구 주제로는 가상 환경 속 캐릭터가 자연스럽고 다양한 움직임을 생성할 수 있도록, 데이터를 효과적으로 활용하는 애니메이션 기술에 초점을 맞추고 있다.

주요 연구 분야는 크게 두 가지로 나뉜다. 하나는 모션 캡처 데이터를 모방하는 물리 기반 시뮬레이션 기법을 활용하여 캐릭터를 제어하는 방법이며, 다른 하나는 확산 모델과 같은 생성형 모델을 통해 주어진 조건을 만족하는 동작을 합성하는 것이다. 특히, 모션 데이터의 다양성과 양이 제한된 상황에서도 학습 효율을 높이는 방법론에 큰 관심을 두고 있으며, 부위별 움직임의 조합이나 다양한 작업에 재사용 가능한 모션 사전 지식의 학습 등을 통해 이를 실현하고자 하였다.

연구 성과로는 AAAI 2022, SIGGRAPH 2023 및 2025, Eurographics 2025 등에 논문을 발표하였고, ICCV 및 CVPR 등 주요 컴퓨터 비전 학회 및 ICRA 2023 Robot Wrestling Challenge에도 공저자로 참여함으로써 애니메이션 기술의 학제 간 확장과 발전에 기여하였다. 또한, Roblox 및 Meta(예정)에서 연구 인턴으로서 기술 개발에 참여한 경험이 있다.

그림 1. 부위별 움직임 조합을 통한 복잡한 인간-물체 상호작용 재현

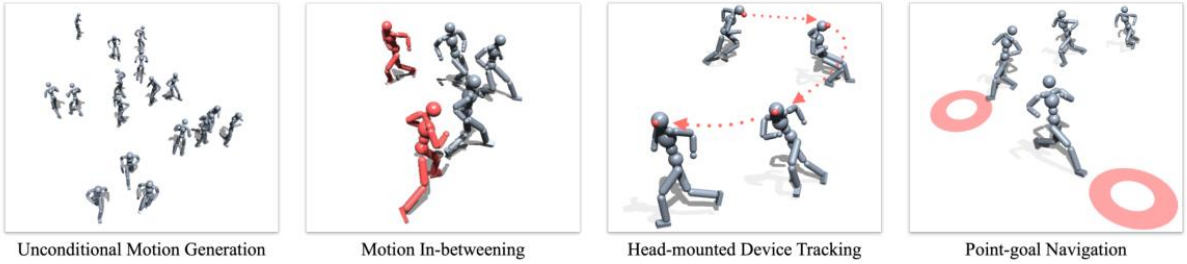


그림 2. 다양한 작업에 재사용 가능한 모션 사전 지식 학습

2025 창해 신진 연구자상 후보

이다원 (Dawon Lee)

dawon.lee@kaist.ac.kr

이다원(Dawon Lee)은 KAIST 문화기술대학원에서 노준용 교수님의 지도 하에 박사학위를 취득하고, 현재 KAIST 박사후연구원으로 재직 중인 컴퓨터 과학자이다. 그는 미디어 콘텐츠 제작과 소비 경험을 기술적으로 향상시키는 것을 핵심 연구 목표로 삼고, 콘텐츠 기술 분야의 다양한 문제를 해결하기 위한 소프트웨어 기반 접근을 연구해 왔다. 특히 컴퓨터 그래픽스, 인간-컴퓨터 상호작용, 멀티미디어 등 여러 분야의 접근을 융합하여, 미디어 콘텐츠 제작 과정의 자동화 및 최적화를 위한 알고리즘을 개발하고 있으며, 사용자의 인지적 특성과 환경적 조건을 반영한 맞춤형 콘텐츠 소비 경험 제공을 위한 방법론 및 인터페이스를 개발하고 있다.

그의 대표적인 연구로는 K-pop 교차편집 영상을 자동으로 생성하는 기술, 사용자가 원하는 길이에 맞게 하이라이트 영상을 자동으로 생성하는 기술, 다양한 디스플레이 환경과 폰트 크기에 적응하는 영상 자막 최적화 표시 기술, 3D 카메라 레이아웃 자동 배치 기술 등이 있다. 연구 성과로 컴퓨터그래픽스와 인간-컴퓨터 상호작용 두 분야 모두에서 세계 최고 권위의 저널 및 학회인 Transactions on Graphics (TOG), SIGGRAGPH Asia, CHI에 최근 3년간 제1저자 및 교신저자로 4편의 논문을 발표했으며, 다수의 국내 및 해외 특허를 등록하고, 기술 이전을 수행한 이력이 있다. 현재는 콘텐츠 제작 자동화 기술의 응용 범위를 확장하기 위한 후속 연구를 진행 중이며, 콘텐츠 제작 도구의 유연성과 사용자 경험 요소를 함께 고려한 기술 설계를 진행하고 있다.

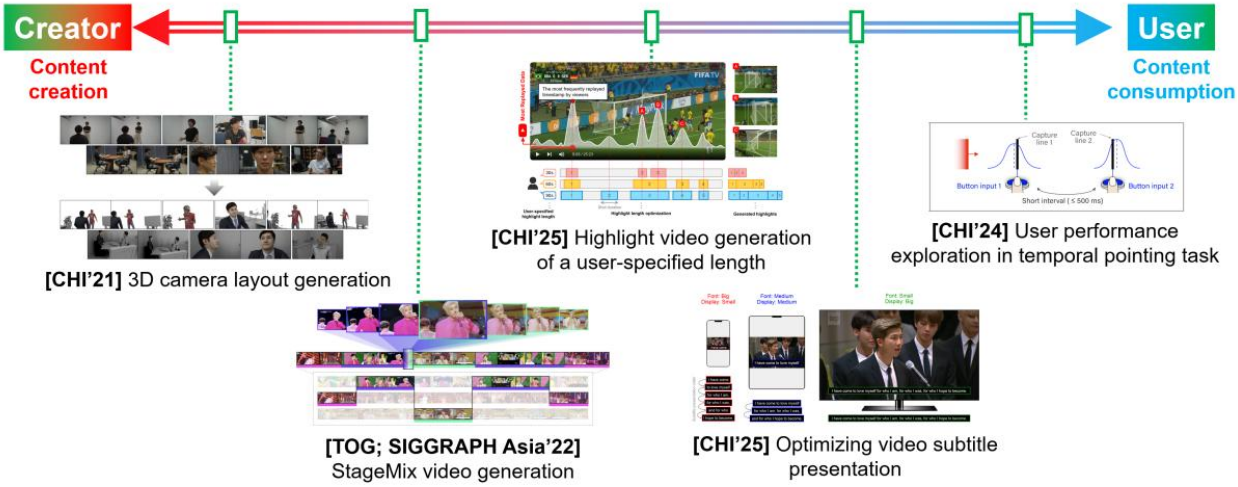


그림 1. 발표한 주요 논문들의 주제 및 적용 맥락에 대한 개요

2025 창해 신진 연구자상 후보

정유철 (Yucheol Jung)

ycjung@postech.ac.kr

정유철(Yucheol Jung)은 3D Cartoon Face Modeling, 3D Shape Deformation, 3D Non-rigid Registration 분야에 관심을 두고 활발히 연구 활동을 이어오고 있는 연구자이다. SIGGRAPH, Eurographics, CVPR 등 주요 그래픽스 및 컴퓨터 비전 학회에서 연구 성과를 발표한 바 있다.

그는 POSTECH 컴퓨터 그래픽스 연구실 출신으로, 대학원생 시절 Microsoft Research Asia와의 공동 연구를 통해 3D Cartoon Face 데이터셋을 구축하였다. 이 과정에서, 카툰에서 자주 나타나는 자연스러운 변형을 컴퓨터 수식으로 모델링하는 방법에 흥미를 가지게 되었다. 이러한 관심은 딥 러닝 기반의 3D 캐리커처 생성 연구와 3D 템플릿 변형 기반의 Non-rigid Registration 기법 연구로 확장되었으며, 특히 카툰 및 캐리커처 처럼 과장되거나 큰 변형이 포함된 3D 형상을 모델링하고 복원하는 데 학술적인 기여를 하였다. 현재는 삼성전자 MX 사업부에서 XR 기술에 활용되는 컴퓨터 비전 소프트웨어의 연구 및 개발을 수행하고 있다.

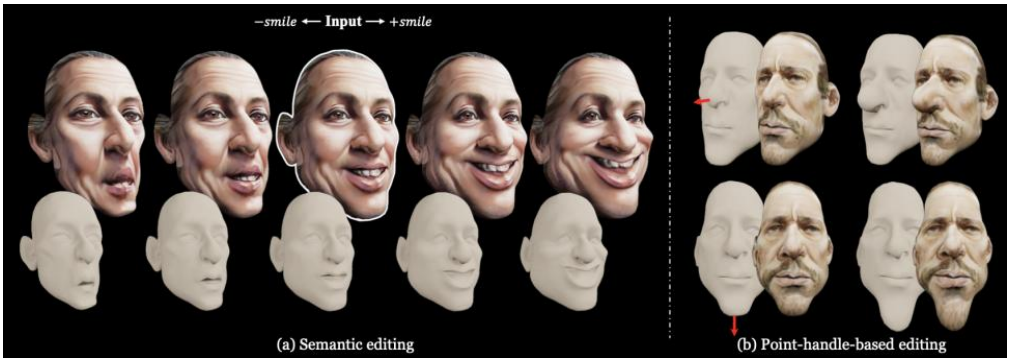


그림 1. 3D 캐치커처 자동생성 연구

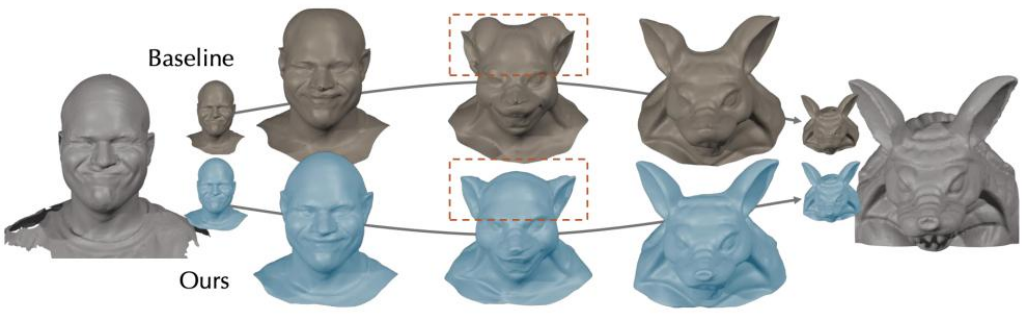


그림 2. 큰 변형이 포함된 3D 형상 간의 Non-rigid Registration 연구

석사 논문상 후보자 발표

2025 석사 논문상 후보

김경민 (KyeongMin Kim)

kgm031189@korea.ac.kr

김경민 학생은 2023년부터 컴퓨터 그래픽스 분야 중 캐릭터 애니메이션에 대해 연구를 시작했다. 자연스러운 애니메이션 합성이 주 연구 목표였으나, 광학식 모션 캡처를 통한 데이터 수집 중 경직된 마커 배치 제약과 이로 인한 노이즈를 문제로 인식하여 해결했다. 해당 기술을 논문으로 작성하여 2024년 ACM Transactions on Graphics에 게재하였으며, 이를 통해 광학식 모션 캡처 시스템의 노이즈 개선과 더불어 모션 캡처 시나리오 및 배우가 장비한 소품에 따라 마커 재배치 및 부위별 추가 배치를 가능하게 했다.

이 과정에서 체득한 기술과 캐릭터 애니메이션 대한 이해를 기반으로 시각적으로 자연스러움과 더불어 물리적으로 타당한 애니메이션 합성 연구에 도전했다. 다만 전체 모션이 아닌, 가상환경 내 중요 부위 중 하나인 손 모션에 집중했다. 강화학습을 통해 저차원의 사용자 입력 신호를 변환하여 물리적 상호작용만으로 물체를 다를 수 있는 손 모션을 합성하는 연구에 참여하여 성공했다. 해당 연구는 ACM SIGGRAPH 2025에 채택되어 2025년 8월 캐나다 밴쿠버에서 발표될 예정이다. 이 과정에서 강화학습을 통한 문제 해결의 발전 가능성을 인지하였으며, 캐릭터 애니메이션 학습을 위한 강화학습 개선 방식에 대해 연구하고 있다.

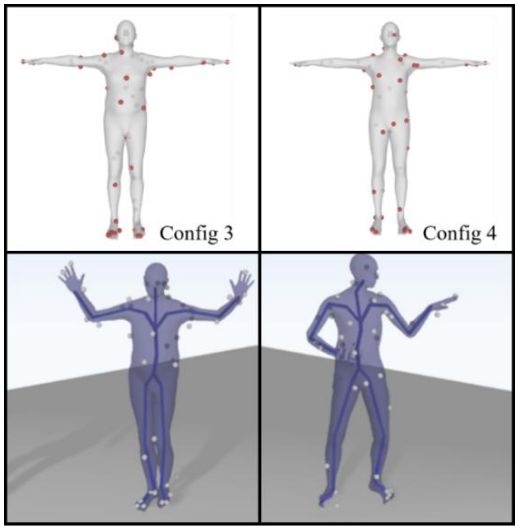


그림 1. 임의 레이아웃 마커의 포즈 복원

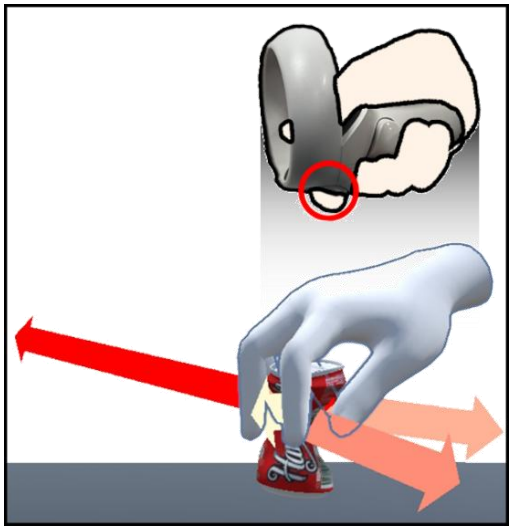


그림 2. 물리적 손 모션 합성

2025 석사 논문상 후보

김민수 (Minsu Kim)

igotaspot426@gmail.com

김민수 학생은 한양대학교 컴퓨터그래픽스&로보틱스 연구실의 석사과정 학생으로, 강화학습을 이용한 캐릭터 애니메이션 생성 및 제어라는 주제로 연구를 진행하고 있다. 2019 년 서울시립대학교 기계정보공학과에서 학사를 취득하였다.

김민수 학생의 주요 연구 내용은 물리환경에서 시뮬레이션되는 휴머노이드 캐릭터의 자연스러운 모션을 학습하고 다양한 태스크를 수행하도록 하는 것이다. 이 연구주제를 축구와 같은 역동적인 움직임이 돋보이는 스포츠 분야에 적용하여 스포츠과학이나 비디오게임에 적용할 수 있음을 보였다. SIGGRAPH 2025에 선정된 논문 PhysicsFC: Learning User-Controlled Skills for a Physics-Based Football Player Controller 에서는 상용 축구게임과 같은 11 vs 11 축구경기를 유저가 컨트롤할 수 있으며 민첩하고 자연스러운 움직임이 가능함을 보였다.

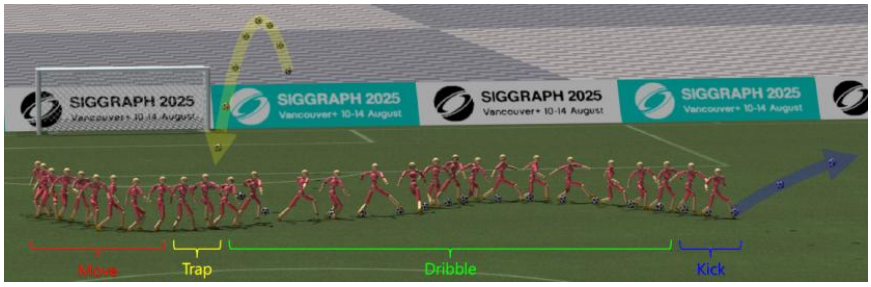


그림 1. 다양한 축구 스킬을 수행하는 모습

그림 2. 11 vs 11 축구 경기 시뮬레이션을 하는 모습.
유저가 직접 컨트롤하고 있다.

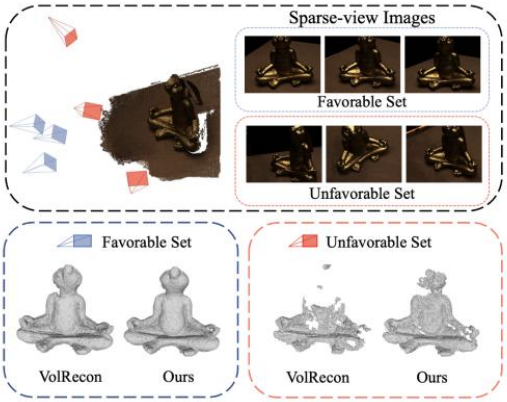
2025 석사 논문상 후보

나영주 (Youngju Na)

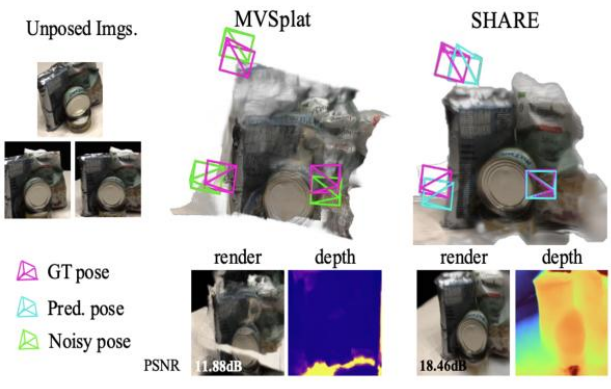
yjna2907@kaist.ac.kr

나영주 학생는 한국과학기술원(KAIST) 전산학부 박사과정에 재학 중이며, 컴퓨터 그래픽스 및 3D 비전 분야, 특히 inverse rendering 및 neural rendering 기술인 Neural Radiance Fields와 3D Gaussian Splatting을 중심으로 연구하고 있다. 전남대학교에서 산업공학과 인공지능을 복수 전공하며 최우수 성적 (GPA: 4.36/4.5)으로 졸업하였고, KAIST에서 전산학 석사 학위(지도교수: 윤성의)를 취득하였다.

석사과정 동안 컴퓨터 그래픽스 및 3D 비전 연구를 수행하여, 최상위 국제 학회인 CVPR 2024에서 "UFORecon: Generalizable Sparse-view Surface Reconstruction from Arbitrary and Unfavorable Sets" 논문을 제1저자로 발표하였다. 본 논문은 2~3장의 이미지 만으로 다양한 환경에서 3차원 표면을 복원하는 방법을 제안했다. 또한, ICIP 2025에서 발표 예정인 "Pose-free 3D Gaussian Splatting via Shape-ray Estimation" 논문을 통해 3D Gaussian splatting 기술을 확장하여 카메라 포즈의 필요성을 없이 feed-forward 방식으로 3차원을 복원하는 방법을 제안하여 확장성 측면에서 기여를 하였다. 또한 비디오로부터 인간, 동물, 로봇 등의 캐릭터 움직임을 추출하고 이를 편집 및 전이하는 2D-to-3D Motion Retargeting 연구를 진행하며, 애니메이션 분야로 연구 영역을 확장하였다. 또한, 후보자는 올해 4월부터 네이버 랩스 Spatial AI 팀에서 연구 인턴으로 참여하여 광범위한 공간에서 inverse-rendering 연구를 수행하고 있다.



(a) UFORecon, CVPR 2024



(b) SHARE, ICIP 2025

2025 석사 논문상 후보

서승원 (SeungWon Seo)

ssw03270@korea.ac.kr

서승원 학생은 고려대학교 컴퓨터학과 석사 과정 3기로 재학 중이다. 현재 IIIXR Lab에서 Embodied Agent를 주제로 연구하고 있으며, 특히 부분 관찰 환경에서의 협력적 계획 수립에 관심을 가지고 있다. 그는 2021년 말 학부 연구생으로 연구를 시작해, 지금까지 이어오고 있다.

첫 번째 연구는 2023년 ACM SIGGRAPH에 DARAM: Dynamic Avatar-Human Motion Remapping Technique for Realistic Virtual Stair Ascending Motions 라는 제목으로 출판되었다. 이 연구는 현실에는 계단이 없지만 가상 환경에는 계단이 존재하는 경우, 사용자가 자연스럽게 계단을 오를 수 있도록 아바타의 모션을 생성하는 환경 불일치 문제를 다룬다. 그는 연구실 밖에서도 다양한 외부 연구 기회를 적극적으로 추구하여, Motion style transfer 연구로 펄어비스 장학금을 수상하였고, 엔씨소프트에서 주관한 NC Fellowship Neural Graphics Track에서 우승을 차지했다. 이를 계기로 엔씨소프트의 Graphics AI Lab에서 여름 인턴십을 수행하며, 제스처 애니메이션 생성에 관한 연구를 진행했다.

이후 다시 연구실로 돌아와 DAMO: A Deep Solver for Arbitrary Marker Configuration in Optical Motion Capture 프로젝트에 참여하였고, 이 연구는 2024년 ACM Transactions on Graphics에 출판되었다. 본 연구는 고정된 마커 구성이 없는 광학식 모션 캡처 환경에서도 인체 포즈를 효율적으로 추정할 수 있는 어텐션 기반 딥러닝 프레임워크를 제안했다.

세 번째 연구이자 그가 주저자로 주도한 첫 번째 연구는 REVECA: Adaptive Planning and Trajectory-based Validation in Cooperative Language Agents Using Information Relevance and Relative Proximity 로, AAAI 2025에 출판되었고 상위 5% 내에 선정되었다. 이 논문은 다중 에이전트 환경에서 협력 가능한 실체화된 에이전트를 주제로 다뤘으며, 이를 위해 정보 관련성 평가, 공간 정보를 활용한 계획 생성, 계획 검증을 다뤘다. 그는 해당 연구를 주제로 2024년 한국연구재단의 석사과정생 연구장려금 지원사업에 선정되었다. 이외에도 ACM CHI 및 ACM Transactions on Graphics의 리뷰어로도 활동했다. 앞으로도 그는 연구를 통해 과학기술 발전에 기여하고, 제가 받아온 여러 기회를 사회에 환원할 수 있는 연구자가 되기를 희망한다.

2025 석사 논문상 후보

신수현 (Suhyun Shin)

shshin9812@postech.ac.kr

신수현은 POSTECH 인공지능대학원 석박통합과정에 재학 중이며, AI, 광학, 컴퓨터 그래픽스를 융합한 초분광 3D 이미징 분야에서 활발한 연구를 수행하고 있다. 특히, 저비용 광학 시스템과 AI 기반 복원 기술을 결합한 새로운 이미징 프레임워크를 제안하며, 세계 최고 권위의 컴퓨터 비전 학회 CVPR에 2024년과 2025년 연속으로 논문을 발표하는 성과를 이루었다. 2024년 논문에서는 RGB 카메라, 프로젝터와 회절 격자 그레이팅을 활용해 파장에 따라 분산되는 구조광을 생성하고, 이를 통해 1mm 깊이 정밀도와 18.8nm 스펙트럼 해상도를 달성하였다. 이어 2025년 논문에서는 조사 패턴의 개수를 8개로 크게 줄여, 동적 장면에서의 실시간 측정을 가능하게 하는 구조광 시스템을 제안하여, 6.6fps 속도, 4mm 깊이 오차, 15.5nm 스펙트럼 해상도를 구현하였다. 전 과정에서 시스템 설계부터 실험, 알고리즘 개발, 성능 검증까지 연구를 주도하였으며, POSTECH, KAUST, Princeton과의 공동 연구를 통해 국제적인 협업 역량도 함께 입증하였다.

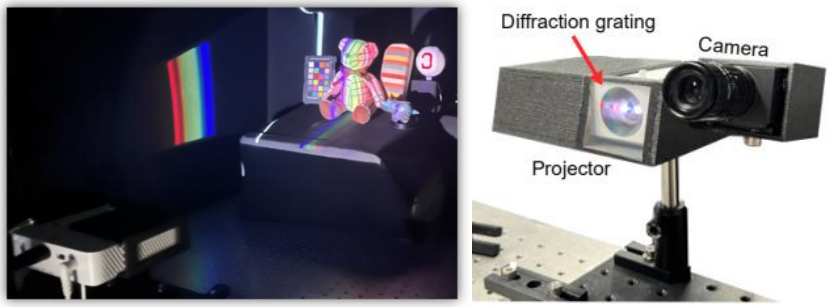


그림 1. CVPR 2024, Dispersed Structured Light for Hyperspectral 3D Imaging 이미징 시스템

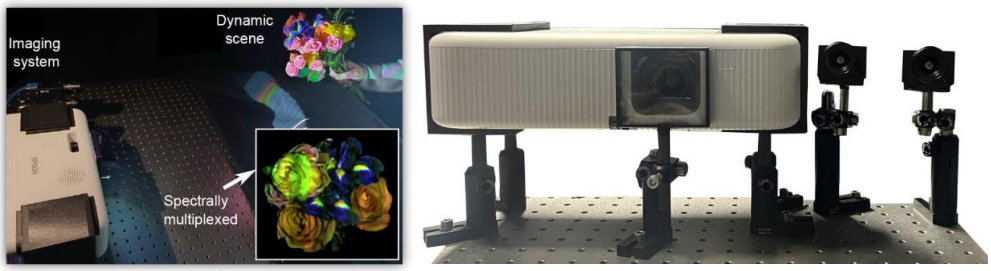


그림 2. CVPR 2025, Dense Dispersed Structured Light for Hyperspectral 3D Imaging of Dynamic Scenes 이미징 시스템

2025 석사 논문상 후보

이수현 (Soohyun Lee)

khtt1222@gmail.com

이수현은 서강대학교 시각컴퓨팅연구실 석사 연구생이다. 서강대학교에서 인공지능 석사 학위를 취득했으며, 정밀한 3D 공간 및 객체 재구성에 관심을 갖고 연구를 수행해왔다. 특히, 희소한 다중 뷰 비디오에서 기하학적으로 일관된 다중 인원 아바타를 재구성하는 "GeoAvatar" 논문을 CVPR 2025에 게재했고, 타원체를 이용한 실용적인 Multi-Center-of-Projection 모델링에 관한 논문을 IEEE Access에 게재했다. 최근에는 빠르고 정확한 3D 재구성 방법에 대해 연구하고 있다.

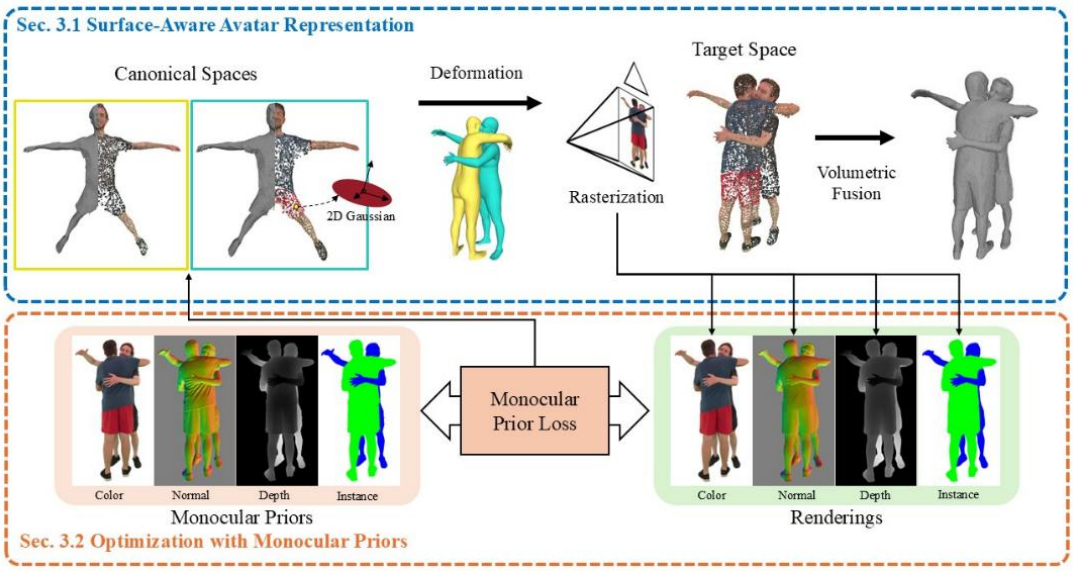


그림 1. GeoAvatar의 파이프라인

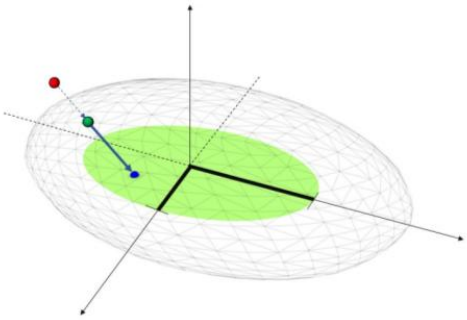


그림 2. 타원을 이용한 Multi-Center-of-Projection 설명도

2025 석사 논문상 후보

이호정 (Hojung Lee)

dearshawn@yonsei.ac.kr

이호정은 연세대학교 컴퓨터과학과 석·박사 통합과정 학생으로, 컴퓨터 그래픽스 연구실에서 이인권 교수님의 지도 하에 통합과정 4학기를 이수 중이다. 2023년 8월 연세대학교에서 컴퓨터과학 학사 학위를 취득하였다.

이호정 학생은 가상현실(VR) 및 인간-컴퓨터 상호작용 (HCI) 분야에서 활발한 연구를 수행하고 있으며, 특히 물리적 공간의 제약을 극복하기 위한 이동 기술, 즉 Redirected Walking(RDW) 기술과 관련된 리셋(reset) 및 리디렉션(redirection) 전략을 주요 주제로 다룬다. 그는 다중 사용자 환경에서 최적의 리셋 방향을 학습하는 강화학습 기반 리셋터(MARR)를 제안하였으며 (그림 1), 사용자의 주변 환경에 따라 다양한 리디렉션 기법 중 최적의 방법을 선택적으로 적용하는 강화학습 기반 Selective Redirection Controller(SRC) 모델을 개발하였다 (그림 2). 또한, 시각과 청각 자극을 결합한 멀티모달 리셋 UI를 고안하여 사용자 경험을 효과적으로 개선하였다. 이러한 연구 결과를 바탕으로 총 3편의 논문을 작성하였고, 모 두 IEEE Transactions on Visualization and Computer Graphics (TVCG) 에 게재되었으며, IEEE VR 및 IEEE ISMAR에서 발표되었다. 이 중 한 편은 2024 IEEE VR에서 Best Paper Award를 수상하였다

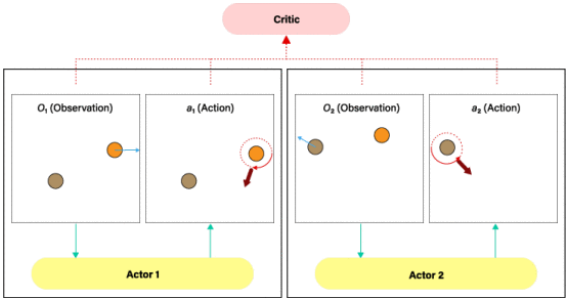


그림 1. 여러 사용자의 리셋 방향을 학습하는 MARR 모델 개요도

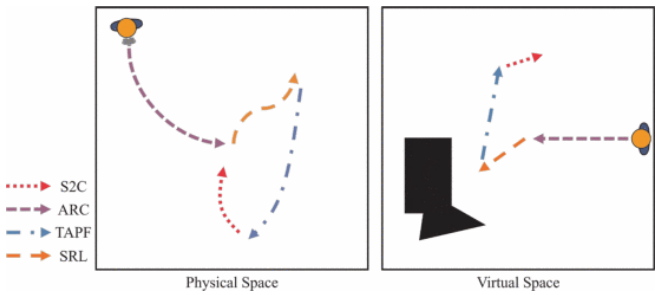


그림 2. 사용자 주변 환경에 따라 최적의 리디렉션 기법을 선택적으로 교차 적용하는 SRC 방법

2025 석사 논문상 후보

임수빈 (Soobin Lim)

dnpcs@korea.ac.kr

임수빈 학생은 2022년 8월 경희대학교 대학원에 입학하여 연구를 시작하였고, 현재는 고려대학교에서 석박사통합과정을 진행 중이다. 초기에는 가상현실 환경에서 계단 오르기 동작의 자연스러움을 개선하는 연구를 수행하였으며, 그 결과물로 "DARAM: Dynamic Avatar-Human Motion Remapping Technique for Realistic Virtual Stair Ascending Motions" 논문을 2023년 SIGGRAPH에서 발표했다. DARAM은 평지 위 사용자와 가상 계단 환경 간의 자세 차이를 분석하여 자연스러운 모션 리매핑을 실현한 기술로, 기존의 정적 계단 한정 연구에서 확장성을 확보한 점이 큰 의의이다. 이후 그는 2023~2024년에는 NC Soft 연구 과제를 통해 성격 기반 NPC 모션 생성 및 텍스트 기반 모션 생성 등 딥러닝 기반 모션 생성 연구를 수행하며 관련 역량을 키웠다. 2024년에 그는 대형 언어 모델(LLM)의 가능성에 주목하여, LLM 다중 에이전트 협력 연구인 "REVECA: Adaptive Planning and Trajectory-based Validation in Cooperative Language Agents using Information Relevance and Relative Proximity"에 공동저자로 참여했다. REVECA는 LLM 에이전트 간 효율적인 협력을 위한 새로운 프레임워크를 제시하며, 2025 AAAI에서 oral paper로 발표되었다. 추후 그는 관심분야인 VR 및 캐릭터 애니메이션에서도 LLM을 활용한 연구를 진행할 계획이다.

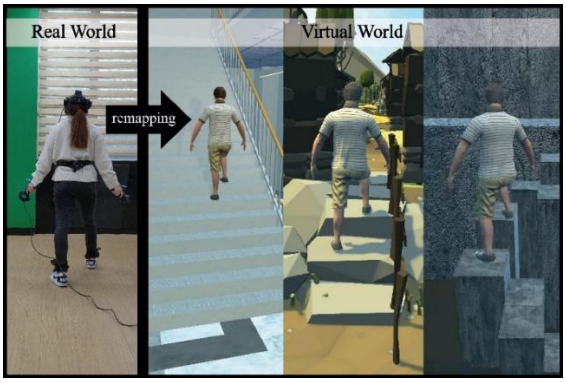


그림 1. DRRAM 예시 그림

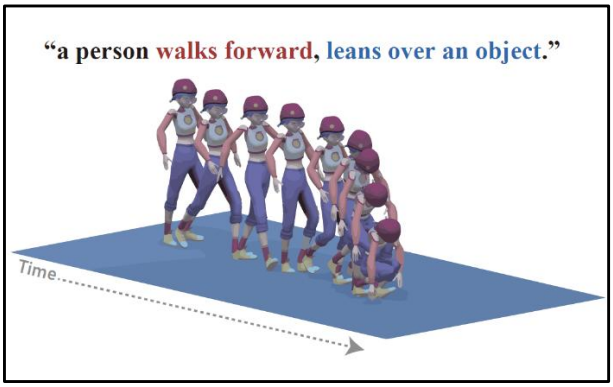


그림 2. Text-to-Motion 예시

2025 석사 논문상 후보

전수민 (Suemin Jeon)

orangeblush@korea.ac.kr

전수민 (Suemin Jeon)은 컴퓨터 그래픽스와 시각화 연구를 하는 고려대학교 박사과정생이다. 특히 몰입형 인터렉티브 시각화 시스템을 중심으로, 복잡한 시각 분석 시스템을 XR 환경에서 누구나 쉽게 저작할 수 있는 플랫폼 개발에 집중해왔다.

주요 연구 성과인 XROps는 별도의 코딩 없이 Web 기반 UI를 통해 XR 시각화 워크플로우를 구성하고 조작할 수 있는 시스템이다. 특히, 노코드 방식의 시각화 편집, 센서 데이터를 포함한 다양한 입력 데이터 처리, 동적 몰입형 분석을 지원함으로써, 환경변화에 능동적으로 대응하는 인터랙티브 분석이 가능하다. 실제 생물학, 스포츠, 의료 등의 다양한 사례를 통해 직관성과 실용성을 입증하였으며, 사용성 평가에서는 비개발자도 몰입형 시각화를 빠르게 구현할 수 있음이 확인되었다. 또한, 최근에는 제한된 하드웨어 환경에서도 과학적 시각화가 가능하도록, 구조 기반의 경량화를 위한 3D Gaussian Splatting 최적화 기법을 제안하였다. 그녀는 이 외에도 과학적 시각화와 사용자 인터페이스 통합, 고속 렌더링 기법에 대한 연구를 지속하며, 실용성과 확장성을 겸비한 인터랙티브 시스템 개발을 통해 학계에 기여하고자 힘쓰고 있다.

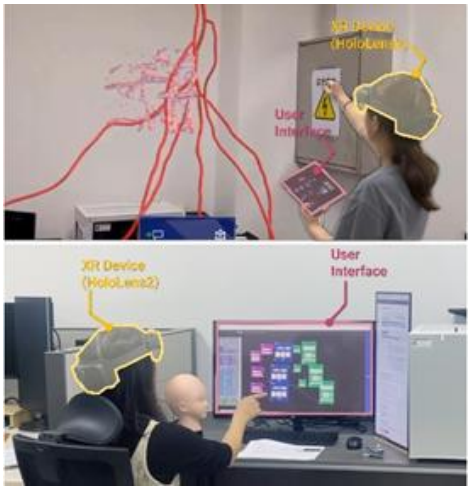


그림 1. XROps 시스템을 활용한
시각화 인터페이스 사용 예시

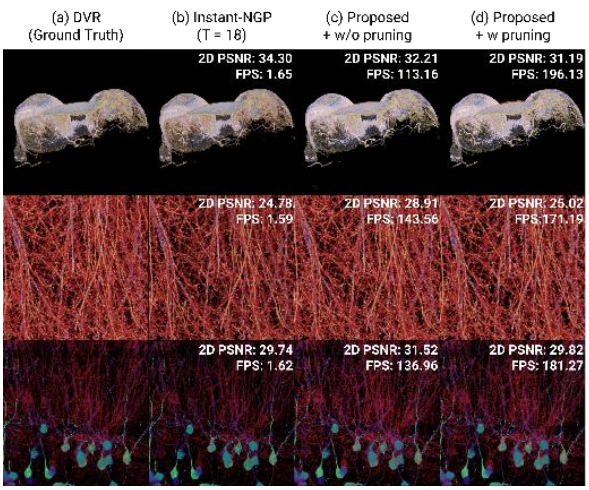


그림 2. 3D Gaussian Splatting을 통한
볼륨 경량화 결과

2025 석사 논문상 후보

황지성 (Jisung Hwang)

4011hjs@kaist.ac.kr

황지성은 KAIST 석사 과정 2학년 재학 중이다. 그는 implicit field에서 메쉬를 추출하는 연구를 진행했다. 최근 신경 암시적 표현(Neural Implicit Representation)이 생성·재구성 분야에서 각광받으면서, implicit field로부터 고품질 메쉬를 추출하는 기술의 중요성이 점점 부각되고 있다. 그러나 기존의 Marching Cubes(MC) 방식은 계단 현상이 발생하고 sharp feature를 보존하지 못하는 한계가 있으며, 이를 보완한 Dual Contouring(DC)·Manifold Dual Contouring(MDC) 등의 기법은 field의 연속적인 도메인 정보를 제대로 활용하지 못하고 제한된 지점에서의 기울기 정보에 의존한다는 근본적인 제약이 있다. 이러한 배경에서 Occupancy-Based Dual Contouring(ODC)를 제안하여, 기울기 정보가 없는 환경에서도 sharp feature를 보존하는 메쉬 추출을 가능하게 하였다.

제안된 방법은 기존 MC, MDC, MISE 대비 우수한 정량적·정성적 성능을 입증하였다. SALAD 실험에서 Chamfer-L1 오차를 SOTA 방법론 대비 평균 6배 감소시켰고, self-intersection을 사실상 방지하였으며, watertightness와 manifoldness를 보장하였다. 또한 학습 과정 수정 없이 기존 neural implicit field에 "drop-in" 방식으로 적용 가능해 대규모 3D 생성 파이프라인에서도 실용적이다. ODC는 SDF 기반 방법론의 기울기 의존성을 제거하면서도 Dual Contouring의 sharp feature 보존 능력을 계승했다는 점에서 의의가 크며, 이를 인정받아 SIGGRAPH Asia 2024에 발표되었다.

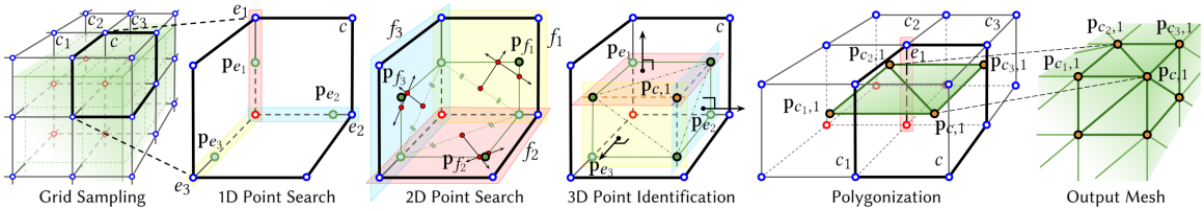


그림 1. OCD의 작동 방식

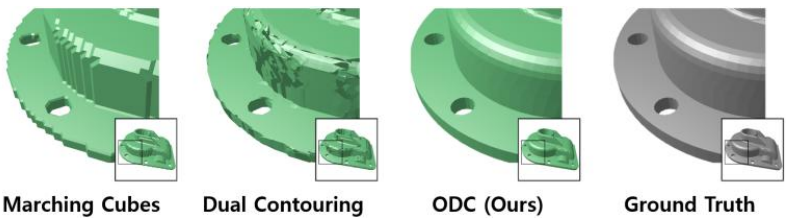


그림 2. 추출된 메쉬 비교. ODC의 결과가 기존 방법 대비 sharp feature를 가장 잘 보존한다.

여 름 학 교

여 름 학 교

AI 기반 컴퓨터 그래픽스/비전 기술은 어떻게 알츠하이머 정밀의료에 활용될 수 있는가?

성준경 교수, 고려대학교

강의 내용

3차원 형상 모델링 또는 처리 기법은 컴퓨터 그래픽스/비전 분야에서 오랫동안 연구되어온 기술로, 다양한 응용 분야에서 활용되고 있다. 알츠하이머병이나 파킨슨병과 같은 퇴행성 뇌질환 진단·예후 예측을 위해 환자의 뇌 MRI 영상분석을 통한 대뇌피질 수축 등 형상 변화 감지와 AI 기반 환자 분류 기술이 중요하며, 3차원 시뮬레이션을 통한 개인 맞춤 치료 과정에서도 컴퓨터 그래픽스 기술이 핵심 역할을 한다. 본 강의에서는 최신 연구 사례를 통해 알츠하이머 정밀의학의 트렌드와 컴퓨터 그래픽스 기술의 활용 방안을 살펴본다.

연사소개

2008-2010	KAIST 연구교수
2010-2012	숭실대학교 조교수
2013-2018	고려대학교 부교수
2018-	고려대학교 바이오공학부 인공지능학과 교수

여 름 학 교

Personalized 3D Human Avatar Generation and Real-Time Animation from a Single Image

문경식 교수, 고려대학교

강의 내용

단일 이미지로부터 개인의 얼굴, 체형, 의상 등 외형 정보를 정밀하게 보존한 3D 아바타를 생성하고, 이를 기반으로 자연스러운 옷의 움직임을 포함한 실시간 애니메이션과 비디오를 제작하는 최신 기술을 다룬다. 표현 학습, 신경 렌더링, 포즈 제어, 텍스처 복원, 생성 기반 비디오 모델 등 폭넓게 소개하며, 단순한 복제 수준을 넘어 개인화된 디지털 휴먼의 활용 가능성을 탐색한다.

연사소개

2011-2015	포항공과대학교 학사
2015-2021	서울대학교 전기정보공학부 박사
2022	서울대학교 박사후연구원
2022-2024	Postdoctoral Research Scientist, Reality Labs Research at Meta
2024	대구경북과학기술원 조교수
2025-	고려대학교 컴퓨터학과 조교수

여 름 학 교

Generative Modeling for Photorealistic 3D Digital Humans

주한별 교수, 서울대학교

강의 내용

사실적인 3차원 디지털 휴먼을 복원하고 생성하는 최신 방법론을 탐구합니다. 2D 영상으로부터 사람의 키포인트를 추출하고 이를 기반으로 3차원 모션과 외형을 복원하며, 생성형 모델을 활용한 디지털 휴먼의 동작 생성 기술을 함께 학습합니다.

연사소개

2007	KAIST 전산학 학사
2009	KAIST 전기 및 전자공학 석사
2009-2012	ETRI 연구원
2018	카네기멜론대학교 박사
2019-2022	Research Scientist, Facebook AI Research (FAIR)
2022-	서울대학교 컴퓨터공학부 조교수

우수 국제 학술대회 논문 발표

이미지/비디오

ICLR 2025

StochSync: Stochastic Diffusion Synchronization for Image Generation in Arbitrary Spaces

여경민, 김재훈, 성민혁

발표자

여경민 (aaaaa@kaist.ac.kr)

요약

We propose a zero-shot method for generating images in arbitrary spaces (e.g., a sphere for 360° panoramas and a mesh surface for texture) using a pretrained image diffusion model. The zero-shot generation of various visual content using a pretrained image diffusion model has been explored mainly in two directions. First, Diffusion Synchronization-performing reverse diffusion processes jointly across different projected spaces while synchronizing them in the target space-generates high-quality outputs when enough conditioning is provided, but it struggles in its absence. Second, Score Distillation Sampling-gradually updating the target space data through gradient descent-results in better coherence but often lacks detail. In this paper, we reveal for the first time the interconnection between these two methods while highlighting their differences. To this end, we propose StochSync, a novel approach that combines the strengths of both, enabling effective performance with weak conditioning. Our experiments demonstrate that StochSync provides the best performance in 360° panorama generation (where image conditioning is not given), outperforming previous finetuning-based methods, and also delivers comparable results in 3D mesh texturing (where depth conditioning is provided) with previous methods.

SIGGRAPH 2025

BrepDiff: Single-stage B-rep Diffusion Model

이민기, 장동수, 클레망 잠봉, 김영민

발표자

이민기 (mingi1019@snu.ac.kr)

요약

The Boundary Representation (B-rep) is a widely used 3D model representation of most consumer products designed with CAD software. However, its highly irregular and sparse set of relationships poses significant challenges for designing a generative model tailored to B-reps. Existing approaches use multi-stage approaches to satisfy the complex constraints sequentially. As a result, the final geometry cannot incorporate user edits due to the non-deterministic dependencies between cascaded stages. In contrast, we propose BrepDiff, a single-stage diffusion model for B-rep generation. We present a masked UV grid representation consisting of structured point samples from faces, serving as input for a diffusion transformer. By introducing an asynchronous and shifted noise schedule, we improve the training signal, enabling the diffusion model to better capture the distribution of UV grids. The explicitness of our masked UV grid representation enables users to intuitively understand and freely design surface geometry without being constrained by topological validity. The interconnectivity can be derived from the face layout, which is later processed into a valid solid volume during post-processing. Our approach achieves performance on par with state-of-the-art cascaded models while offering complex and diverse manipulations of geometry and topology, such as shape completion, merging, and interpolation.

SIGGRAPH 2025

Elevating 3D Models: High-Quality Texture and Geometry Refinement from a Low-Quality Model

류누리, 원지윤, 손주은, 공민수, 이주행, 조성현

발표자

류누리 (ryunuri@postech.ac.kr)

요약

High-quality 3D assets are essential for various applications in computer graphics and 3D vision but remain scarce due to significant acquisition costs. To address this shortage, we introduce Elevate3D, a novel framework that transforms readily accessible low-quality 3D assets into higher quality. At the core of Elevate3D is HFS-SDEdit, a specialized texture enhancement method that significantly improves texture quality while preserving the appearance and geometry while fixing its degradations. Furthermore, Elevate3D operates in a view-by-view manner, alternating between texture and geometry refinement. Unlike previous methods that have largely overlooked geometry refinement, our framework leverages geometric cues from images refined with HFS-SDEdit by employing state-of-the-art monocular geometry predictors. This approach ensures detailed and accurate geometry that aligns seamlessly with the enhanced texture. Elevate3D outperforms recent competitors by achieving state-of-the-art quality in 3D model refinement, effectively addressing the scarcity of high-quality open-source 3D assets.

SIGGRAPH 2025

DC-VSR: Spatially and Temporally Consistent Video Super-Resolution with Video Diffusion Prior

한장혁, 심규진, 김건웅, 이현승, 최규하, 한영석, 조성현

발표자

심규진 (sgj0402@postech.ac.kr)

요약

Video super-resolution (VSR) aims to reconstruct a high-resolution (HR) video from a low-resolution (LR) counterpart. Achieving successful VSR requires producing realistic HR details and ensuring both spatial and temporal consistency. To restore realistic details, diffusion-based VSR approaches have recently been proposed. However, the inherent randomness of diffusion, combined with their tile-based approach, often leads to spatio-temporal inconsistencies. In this paper, we propose DC-VSR, a novel VSR approach to produce spatially and temporally consistent VSR results with realistic textures. To achieve spatial and temporal consistency, DC-VSR adopts a novel Spatial Attention Propagation (SAP) scheme and a Temporal Attention Propagation (TAP) scheme that propagate information across spatio-temporal tiles based on the self-attention mechanism. To enhance high-frequency details, we also introduce Detail-Suppression Self-Attention Guidance (DSSAG), a novel diffusion guidance scheme. Comprehensive experiments demonstrate that DC-VSR achieves spatially and temporally consistent, high-quality VSR results, outperforming previous approaches.

CVPR2025

Dense Dispersed Structured Light for Hyperspectral 3D Imaging of Dynamic Scenes

신수현, 윤승우, Ryota Maeda, 백승환

발표자

신수현 (shshin9812@postech.ac.kr)

요약

Hyperspectral 3D imaging captures both depth maps and hyperspectral images, enabling comprehensive geometric and material analysis. Recent methods achieve high spectral and depth accuracy; however, they require long acquisition times—often over several minutes—or rely on large, expensive systems, restricting their use to static scenes. We present Dense Dispersed Structured Light (DDSL), an accurate hyperspectral 3D imaging method for dynamic scenes that utilizes stereo RGB cameras and an RGB projector equipped with an affordable diffraction grating film. We design spectrally multiplexed DDSL patterns that significantly reduce the number of required projector patterns, thereby accelerating acquisition speed. Additionally, we formulate an image formation model and a reconstruction method to estimate a hyperspectral image and depth map from captured stereo images. As the first practical and accurate hyperspectral 3D imaging method for dynamic scenes, we experimentally demonstrate that DDSL achieves a spectral resolution of 15.5 nm full width at half maximum (FWHM), a depth error of 4 mm, and a frame rate of 6.6 fps.

CVPR2025

Differentiable Inverse Rendering with Interpretable Basis BRDFs

정훈규, 최석준, 백승환

발표자

정훈규 (hgchung@postech.ac.kr)

요약

Inverse rendering seeks to reconstruct both geometry and spatially varying BRDFs (SVBRDFs) from captured images. To address the inherent ill-posedness of inverse rendering, basis BRDF representations are commonly used, modeling SVBRDFs as spatially varying blends of a set of basis BRDFs. However, existing methods often yield basis BRDFs that lack intuitive separation and have limited scalability to scenes of varying complexity. In this paper, we introduce a differentiable inverse rendering method that produces interpretable basis BRDFs. Our approach models a scene using 2D Gaussians, where the reflectance of each Gaussian is defined by a weighted blend of basis BRDFs. We efficiently render an image from the 2D Gaussians and basis BRDFs using differentiable rasterization and impose a rendering loss with the input images. During this analysis-by-synthesis optimization process of differentiable inverse rendering, we dynamically adjust the number of basis BRDFs to fit the target scene while encouraging sparsity in the basis weights. This ensures that the reflectance of each Gaussian is represented by only a few basis BRDFs. This approach enables the reconstruction of accurate geometry and interpretable basis BRDFs that are spatially separated. Consequently, the resulting scene representation, comprising basis BRDFs and 2D Gaussians, supports physically-based novel-view relighting and intuitive scene editing.

우수 국제 학술대회 논문 발표

캐릭터 컨트롤

SIGGRAPH 2025

PhysicsFC: Learning User-Controlled Skills for a Physics-Based Football Player Controller

김민수, 정은호, 이윤상

발표자

김민수 (igotaspot426@gmail.com)

요약

We propose PhysicsFC, a method for controlling physically simulated football player characters to perform a variety of football skills–such as dribbling, trapping, moving, and kicking–based on user input, while seamlessly transitioning between these skills. Our skill-specific policies, which generate latent variables for each football skill, are trained using an existing physics-based motion embedding model that serves as a foundation for reproducing football motions. Key features include a tailored reward design for the Dribble policy, a two-phase reward structure combined with projectile dynamics-based initialization for the Trap policy, and a Data-Embedded Goal-Conditioned Latent Guidance (DEGCL) method for the Move policy. Using the trained skill policies, the proposed football player finite state machine (PhysicsFC FSM) allows users to interactively control the character. To ensure smooth and agile transitions between skill policies, as defined in the FSM, we introduce the Skill Transition-Based Initialization (STI), which is applied during the training of each skill policy. We develop several interactive scenarios to showcase PhysicsFC's effectiveness, including competitive trapping and dribbling, give-and-go plays, and 11v11 football games, where multiple PhysicsFC agents produce natural and controllable physics-based football player behaviors. Quantitative evaluations further validate the performance of individual skill policies and the transitions between them, using the presented metrics and experimental designs.

SIGGRAPH 2025

PLT: Part-wise Latent Tokens as Adaptable Motion Priors for Physically Simulated Character

배진석, 이영환, 임동근, 김영민

발표자

배진석 (capoo4938@gmail.com)

요약

Physically simulated characters can learn highly natural full-body motion guided by motion capture datasets. However, the range of motion is limited to the existing high-quality datasets, and cannot effectively adapt to challenging scenarios. We propose a novel policy architecture that learns part-wise motion skills, where individual parts can be separately extended and combined for unobserved settings. Our method employs a set of part-specific codebooks, which robustly capture motion dynamics without catastrophic collapse or forgetting. This structured decomposition allows intuitive control over the character's behavior and dynamic exploration for a novel combination of part-wise motion. We further incorporate a refinement network compensating for subtle discrepancies in the disjoint discrete tokens, thus improving motion quality and stability. Our extensive evaluations show that our part-wise latent token achieves superior performance in imitating motions, even those from unseen distribution. We also validate our method in challenging tasks, including body tracking, navigation on complex terrains, and point-goal navigation with damaged body parts. Finally, we introduce a part-wise expansion of motion priors, where the physically simulated character incrementally adapts partial motion and produces unique combinations of whole-body motion, significantly diversifying motions.

CVPR 2025

AnyMoLe: Any Character Motion In-betweening Leveraging Video Diffusion Models

윤관, 홍석현, 김채린, 노준용

발표자

윤관 (yunandy@kaist.ac.kr)

요약

Despite recent advancements in learning-based motion in-betweening, a key limitation has been overlooked: the requirement for character-specific datasets. In this work, we introduce AnyMoLe, a novel method that addresses this limitation by leveraging video diffusion models to generate motion in-between frames for arbitrary characters without external data. Our approach employs a two-stage frame generation process to enhance contextual understanding. Furthermore, to bridge the domain gap between real-world and rendered character animations, we introduce ICAdapt, a fine-tuning technique for video diffusion models. Additionally, we propose a motion-video mimicking' optimization technique, enabling seamless motion generation for characters with arbitrary joint structures using 2D and 3D-aware features. AnyMoLe significantly reduces data dependency while generating smooth and realistic transitions, making it applicable to a wide range of motion in-betweening tasks.

CVPR 2025

SALAD: Skeleton-aware Latent Diffusion for Text-driven Motion Generation and Editing

홍석현, 김채린, 윤세린, 남정현, 차시헌, 노준용

발표자

홍석현 (ghd3079@kaist.ac.kr)

요약

Text-driven motion generation has advanced significantly with the rise of denoising diffusion models. However, previous methods often oversimplify representations for the skeletal joints, temporal frames, and textual words, limiting their ability to fully capture the information within each modality and their interactions. Moreover, when using pre-trained models for downstream tasks, such as editing, they typically require additional efforts, including manual interventions, optimization, or fine-tuning. In this paper, we introduce a skeleton-aware latent diffusion (SALAD), a model that explicitly captures the intricate inter-relationships between joints, frames, and words. Furthermore, by leveraging cross-attention maps produced during the generation process, we enable the attention-based zero-shot text-driven motion editing using a pre-trained SALAD model, requiring no additional user input beyond text prompts. Our approach significantly outperforms previous methods in terms of text-motion alignment without compromising generation quality, and demonstrates practical versatility by providing diverse editing capabilities beyond generation. Code is available at project page.

SIGGRAPH 2025

ViSA: Physics-based Virtual Stunt Actors for Ballistic Stunts

김민석, 서원정, 이성희, 원정담

발표자

김민석 (minseok@imo.snu.ac.kr)

요약

We introduce ViSA (Virtual Stunt Actors), an interactive animation system designed to create realistic ballistic stunt actions frequently seen in filmmaking and TV production. By providing spatial constraints suitable for the desired stunt scene, our system generates physically plausible motions satisfying the given constraints. The problem is formulated as a deep reinforcement learning task, incorporating a novel state and action spaces, as well as straightforward yet effective rewards for ballistic stunt actions. Users can receive a fast response within several minutes and continue to choreograph complex stunt scenes in an interactive manner. We demonstrate ballistic stunt scenes resembling those in various films and TV dramas, such as traffic accidents, falling down stairs, and falls from buildings. The effectiveness of the technical components and design choices in our system is demonstrated through extensive comparisons, analyses, and ablation studies.

SIGGRAPH 2025

MAGNET: Muscle Activation Generation Networks for Diverse Human Movement

박정남, 정의균, 이제희, 원정담

발표자

정의균 (jek5224@imo.snu.ac.kr)

요약

We introduce MAGNET (Muscle Activation Generation Networks), a scalable framework for reconstructing full-body muscle activations across diverse human movements. Our approach employs musculoskeletal simulation with a novel two-level controller architecture trained using three-stage learning methods. Additionally, we develop distilled models tailored for solving downstream tasks or generating real-time muscle activations, even on edge devices. The efficacy of our framework is demonstrated through examples of daily life and challenging behaviors, as well as comprehensive evaluations.

우수 국제 학술대회 논문 발표

메쉬/VR/AR

SIGGRAPH Asia 2024

Occupancy-Based Dual Contouring

황지성, 성민혁

발표자

황지성 (4011hjs@kaist.ac.kr)

요약

We introduce a dual contouring method that provides state-of-the-art performance for occupancy functions while achieving computation times of a few seconds. Our method is learning-free and carefully designed to maximize the use of GPU parallelization. The recent surge of implicit neural representations has led to significant attention to occupancy fields, resulting in a wide range of 3D reconstruction and generation methods based on them. However, the outputs of such methods have been underestimated due to the bottleneck in converting the resulting occupancy function to a mesh. Marching Cubes tends to produce staircase-like artifacts, and most subsequent works focusing on exploiting signed distance functions as input also yield suboptimal results for occupancy functions. Based on Manifold Dual Contouring (MDC), we propose Occupancy-based Dual Contouring(ODC), which mainly modifies the computation of grid edge points (1D points) and grid cell points (3D points) to not use any distance information. We introduce auxiliary 2D points that are used to compute local surface normals along with the 1D points, helping identify 3D points via the quadric error function. To search the 1D, 2D, and 3D points, we develop fast algorithms that are parallelizable across all grid edges, faces, and cells. Our experiments with several 3D neural generative models and a 3D mesh dataset demonstrate that our method achieves the best fidelity compared to prior works.

SIGGRAPH 2025

ForceGrip: Reference-Free Curriculum Learning for Realistic Grip Force Control in VR Hand Manipulation

한동현, 김병민, 이로운, 김경민, 황효석, 강형엽

발표자

한동현 (hand32@khu.ac.kr)

요약

Realistic hand manipulation is a key component of immersive virtual reality (VR), yet existing methods often rely on kinematic approaches or motion-capture datasets that omit crucial physical attributes such as contact forces and finger torques. Consequently, these approaches prioritize tight, one-size-fits-all grips rather than reflecting users' intended force levels. We present ForceGrip, a deep learning agent that synthesizes realistic hand manipulation motions, faithfully reflecting the user's grip force intention. Instead of mimicking predefined motion datasets, ForceGrip uses generated training scenarios—randomizing object shapes, wrist movements, and trigger input flows—to challenge the agent with a broad spectrum of physical interactions. To effectively learn from these complex tasks, we employ a three-phase curriculum learning framework comprising Finger Positioning, Intention Adaptation, and Dynamic Stabilization. This progressive strategy ensures stable hand-object contact, adaptive force control based on user inputs, and robust handling under dynamic conditions. Additionally, a proximity reward function enhances natural finger motions and accelerates training convergence. Quantitative and qualitative evaluations reveal ForceGrip's superior force controllability and plausibility compared to state-of-the-art methods.

AAAI2025

REVECA: adaptive planning and trajectory-based validation in cooperative language agents using information relevance and relative proximity

서승원, 노성래, 이준혁, 이원희, 강형엽

발표자

노성래 (rhosunr99@korea.ac.kr)

요약

We address the challenge of multi-agent cooperation, where agents achieve a common goal by cooperating with decentralized agents under complex partial observations. Existing cooperative agent systems often struggle with efficiently processing continuously accumulating information, managing globally suboptimal planning due to lack of consideration of collaborators, and addressing false planning caused by environmental changes introduced by other collaborators. To overcome these challenges, we propose the RElevance, Proximity, and Validation-Enhanced Cooperative Language Agent (REVECA), a novel cognitive architecture powered by GPT-4o-mini. REVECA enables efficient memory management, optimal planning, and cost-effective prevention of false planning by leveraging Relevance Estimation, Adaptive Planning, and Trajectory-based Validation. Extensive experimental results demonstrate REVECA's superiority over existing methods across various benchmarks, while a user study reveals its potential for achieving trustworthy human-AI cooperation.

TVCG 2025

Integrating User Input in Automated Object Placement for Augmented Reality

Jalal Safari Bazargani, Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi

발표자

Jalal Safari Bazargani (j.safarib@sju.ac.kr)

요약

Object placement in Augmented Reality (AR) is crucial for creating immersive and functional experiences. However, a critical research gap exists in combining user input with efficient automated placement, particularly in understanding spatial relationships and optimal placement. This study addresses this gap by presenting a novel object placement pipeline for AR applications that balances automation with user-directed placement. The pipeline employs entity recognition, object detection, depth estimation along with spawn area allocation to create a placement system. We compared our proposed method against manual placement in a comprehensive evaluation involving 50 participants. The evaluation included user experience questionnaires, a comparative study of task performance, and post-task interviews. Results indicate that our pipeline significantly reduces task completion time while maintaining comparable accuracy to manual placement. The UEQ-S and TENS scores revealed high user satisfaction. While manual placement offered more direct control, our method provided a more streamlined, efficient experience. This study contributes to the field of object placement in AR by demonstrating the potential of automated systems to enhance user experience and task efficiency.

논문 발표

가상/증강현실

정렬된 여러 공간의 가시성을 조정하는 다수 사용자 기반 텔레프레즌스 시스템*

김태희, 신지훈, 김혜심, 장혁진, 강지호, 이성희
한국과학기술원
{hayleyy321, jihun.shin, hyedeep, jang5s, jhkang0408, sunghee.lee}@kaist.ac.kr

Visibility Modulation of Aligned Spaces for Multi-User Telepresence

Taehei Kim, Jihun Shin, Hyeshim Kim, Hyuckjin Jang, Jiho Kang, Sung-Hee Lee
KAIST

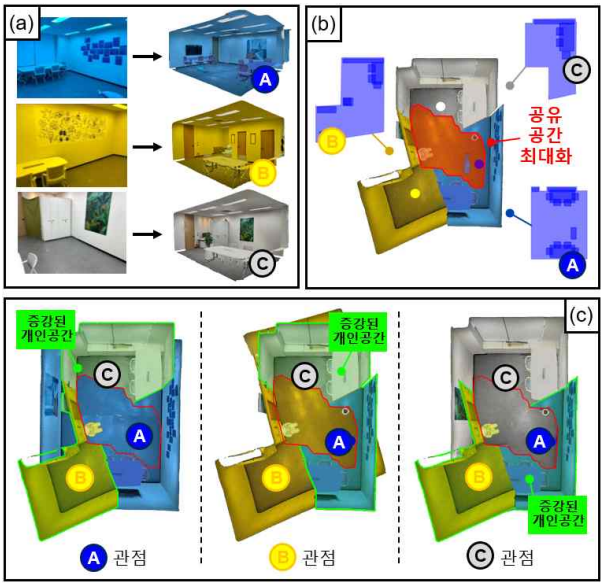


그림 1: 시스템 개요 (a) 사용자 A, B, C의 물리적 실공간의 디지털 카피 (b) 사용자 A, B, C의 공간에서 2D 평면도를 추출하여 공유 공간을 최대화하는 방식으로 정렬시킨 모습 (c) 사용자들의 위치에 따라 각자의 개인 공간의 보여지는 범위가 결정된 동일한 순간의 사용자 A, B, C 관점

요약

본 연구는 사용자 위치에 따라 여러 사용자 방의 보여지는 범위를 동적으로 조정하는 다중 사용자 혼합현실(MR) 텔레프레즌스 시스템을 제안한다. 본 방법은 먼저 공유 공간을 최대화하기 위해 방들의 정렬을 최적화한다. 여기서 공유 공간이란, 다른 사용자 및 실제 혹은 가상 오브젝트와 상호작용할 수 있는 공통 영역이다. 이후 공유 공간 외 공간, 즉 개인 공간을 시각화하여 개인 공간에서 일어나는 활동도 전달되게 한다. 특히, 방의 겹침을 해결하기 위해 사용자의 위치에 따라 개인 공간의 보여지는 범위를 동적으로 조정한다. 본 논문은 이를 활용한 게임 데모에 관한 설명을 포함하고 있다.

* ISMAR 2024 데모 발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)의 데모 세션에 게재되었음.
* 본 연구는 RAPA의 메타버스랩 프로젝트와 한국연구재단(NRF)의 지원으로 수행되었음.

1. 서론

텔레프레즌스 연구 분야는 원격의 사용자들이 “함께 존재하는” 느낌을 실현하기 위해 여러 시스템을 제안해 왔다. 초기 연구에서는 원격 가상 공간이 마치 물리적 공간의 연장처럼 보이도록 구현하거나, [6]과 같이 원격 사용자를 로컬 공간에 아바타 형태로 초대하는 방식이 있었다. 하지만 대부분의 기존 연구는 공간 간 차이를 고려하지 않거나, 두 공간만을 다루는 경우가 많았다. 본 연구는 두 개 이상의 서로 다른 공간을 다루고자 한다. 특히, 각 사용자의 개인 공간을 시각적으로 공유함으로써 텔레프레즌스 경험이 향상될 수 있다는 초기 연구[7, 1]에 영감을 받았으며, 가상 콘텐츠를 물리 공간에 통합할 수 있는 혼합현실의 강점을 적극적으로 활용한다[3]. 이러한 관점에서, 본 시스템은 여러 공간을 정렬하여 상호 작용과 소통이 가능한 공간을 극대화 하고자 한다. 이를 위해 최대 공유 공간을 확보하고, 사용자 위치에 따라 비공유(개인) 공간의 보여지는 범위를 실시간으로 조정한다. 이 과정은 다른 공간에 있는 사용자가 마치 자신의 공간으로 “들어오는” 듯한 독특한 경험을 제공하며, 동시에 개인 공간도 함께 시각화되어 전체적으로 하나로 연결된 연속적인 공간감을 만들어 낸다.

2. 관련 연구

몰입형 3D 텔레프레즌스를 실현하기 위해, 연구자들은 공유 공간 생성[5], 올바른 공간 맥락 전달을 위한 사용자 재배치[8], 부분 정렬[2], 그리고 모션 리타게팅[4] 등 다양한 방법들을 개발해왔다. 이러한 기존 연구들을 바탕으로, 본 연구는 단순히 원격의 사용자를 아바타의 형태로 로컬 공간에 증강하는 수준을 넘어서, 사용자들의 실공간 디지털 카피를 결합시키는 새로운 시스템을 제안한다.

3. 시스템 구현

3.1. 공유 공간 최적화 (Space Optimization)

공유 공간 최적화의 목표는 공유 공간를 최대로 확보할 수 있는 방 정렬을 찾는 것이다. 정렬 과정은 크게 세 단계로 구성되며, 빈 공간(Free Space) 정렬, 객체 증

강, 부드러운 경계 처리로 나뉜다. 먼저, 2D 평면도를 이용하여 각 사용자가 자유롭게 이동할 수 있는 영역을 추출한다. 이후 이 영역들이 최대한 겹쳐지도록 정렬하여 공유 공간을 최대한 확보한다. 다음 단계에서는 공유 공간 영역 내에 위치했던 가상 객체를 증강한다. 마지막으로, 공유 공간의 가장자리 부분을 부드럽게 처리하는 과정을 거친다. 공간 정렬이 완료된 후, 사용자의 위치를 기준으로 비공유 공간의 가시성을 결정한다.

가시성을 결정하는 로직은 사용자의 현재 위치 주변에 존재하는 비공유 공간이 최대한 넓게 보이도록 한다. 이 전략은 사용자의 행동이 어떠한 공간적 맥락에서 일어나고 있는지를 명확하게 전달하기 위함이다. 그 결과, 원격 사용자의 아바타 주변 원격 공간의 일부가 증강되어 로컬 공간과 연결되는 모습을 만들어 낸다. 본 연구에서 제안하는 동적 가시성 조정 방식은 사용자가 공유 공간이든 아니든 관계없이 방 안에서 자유롭게 이동할 수 있도록 해준다. 같은 원리로, 다른 사용자의 전체 공간도 관찰할 수 있게 된다.

3.2. 혼합현실 시스템 (Mixed Reality System)

제안한 방법의 가능성을 검증하기 위해 혼합현실 프로토타입 시스템을 개발했다. 세 명의 사용자가 참여하는 텔레프레즌스 시나리오를 기반으로, 실제 방의 3D 스캔을 입력으로 받아 Meta Quest 3와 Unity 엔진, Meta XR Core SDK를 사용해 구현했다. 데모 설계를 위해 Netcode 네트워크를 활용하여 사용자들이 경험하는 환경을 동기화했다.

4. 데모

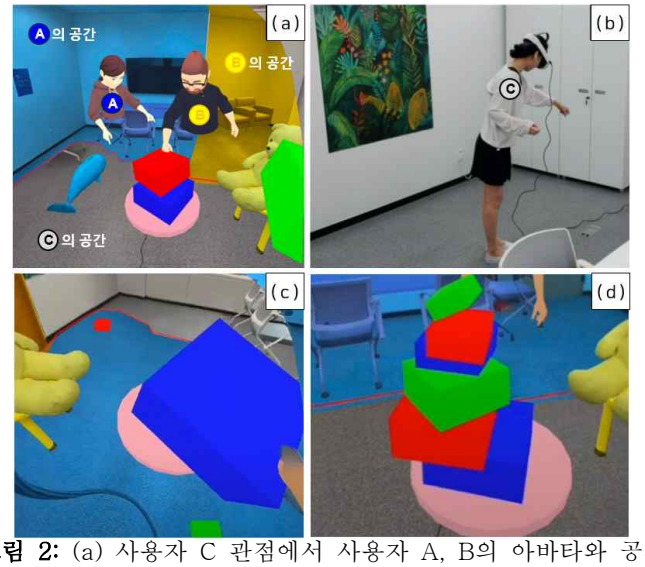


그림 2: (a) 사용자 C 관점에서 사용자 A, B의 아바타와 공간이 함께 보이는 모습 (b) 실제 공간에서 시스템 체험 중인 사용자 C의 모습 (c) 조각이 공간 내에 흩어져 있는 초기 모습 (d) 공간 내의 조각을 협력하여 모두 찾아 탑을 완성한 모습

이 시스템의 데모는 세 명의 사용자가 각기 다른 실제 방에서 접속해 참여한다. 접속 시, 세 사용자 모두 자유롭게 이동할 수 있는 공유 공간이 붉은 선으로 표시된다. 사용자는 자신의 방뿐 아니라 다른 두 사용자의 방도 보게 되며, 움직이는 사용자의 위치에 따라 증강되어 시각화되는 비공유 공간의 범위도 동적으로 변한다.

모든 사용자가 접속하면 첫 번째 단계인 '방 소개'가 시작된다. 각 사용자는 자신의 방을 돌아다니며 내부 구조를 설명하고, 이를 통해 서로의 공간을 이해하게 된다. 이후 두 번째 단계인 '탑 쌓기' 게임이 진행된다. 공유 공간의 조각은 모두에게 보이지만, 개인 공간의 조각은 가시성에 따라 보이지 않을 수 있다. 사용자는 협력해 조각을 모아 탑을 완성해야 하며, 이 과정을 통해 공간들이 하나로 연결된 듯한 경험을 하게 된다.

참고문헌

[1] S. Gibbs, C. Arapis, and C. Breiteneder. Teleport - towards immersive copresence. Multimedia Syst., 7:214–221, 05 1999. doi: 10.1007/s005300050123

[2] J. E. S. Grønbæk, K. Pfeuffer, E. Velloso, M. Astrup, M. I. S. Pedersen, M. Kjær, G. Leiva, and H. Gellersen. Partially blended realities: Aligning dissimilar spaces for distributed mixed reality meetings. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581515

[3] J. Hartmann, C. Holz, E. Ofek, and A. D. Wilson. Realitycheck: Blending virtual environments with situated physical reality. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300577

[4] J. Kang, D. Yang, T. Kim, Y. Lee, and S.-H. Lee. Real-time retargeting of deictic motion to virtual avatars for augmented reality telepresence. In 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 885–893, 10 2023. doi: 10.1109/ISMAR59233.2023.
00104

[5] M. Keshavarzi, M. Zollhoefer, A. Y. Yang, P. Peluse, and L. Caldas. Mutual scene synthesis for mixed reality telepresence, 2022.

[6] S. Orts-Escolano et al. Holoportation: Virtual 3d teleportation in realtime. In Proc. 29th Annual Symposium on User Interface Software and Technology, UIST '16, pp. 741–754. ACM, 2016. doi: 10.1145/2984511.2984517

[7] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '98, p. 179–188. Association for Computing Machinery, New York, NY, USA, 1998. doi: 10.1145/280814.280861

[8] D. Yang, J. Kang, T. Kim, and S.-H. Lee. Visual guidance for user placement in avatar-mediated telepresence between dissimilar spaces. IEEE Transactions on Visualization and Computer Graphics, p. 1–14, 2024. doi: 10.1109/tvcg.2024.3354256

시청각 자극 기반 가상현실 사용자를 위한 방향전환보행 리셋 인터페이스 비교 분석

이호정[0,1], 김현정[1,2], 이인권[1]
연세대학교 컴퓨터과학과[1], 취리히 연방 공과대학교 게임기술센터[2]
dearshawn@yonsei.ac.kr, hcihjkim@gmail.com, iklee@yonsei.ac.kr

Multimodal Turn in Place: A Comparative Analysis of Visual and Auditory Reset UIs in Redirected Walking

Ho Jung Lee[0,1], Hyunjeong Kim[1,2], In-Kwon Lee[1]
Dept. of Computer Science, Yonsei Univeristy[1], Game Technology Center, ETH Zürich[2]

요약

방향전환보행(Redirected Walking; RDW)의 리셋은 가상현실을 체험하는 사용자가 제한된 현실 공간에서도 연속적이고 충돌 없는 보행 경험을 가능하게 하지만, 잦은 리셋은 사용자의 몰입감을 저하시킨다. 이러한 문제를 해결하기 위해 최적의 리셋 방향을 안내하는 다양한 리셋 기술과 시각적 리셋 사용자 인터페이스(UI)가 제안되어 왔지만, 그 효과성은 아직 충분히 검증되지 않았다. 본 연구에서는 RDW를 다년간 연구한 전문가들을 대상으로 인터뷰를 진행하여, 기존 시각적 리셋 UI가 사용자가 제때 인식하지 못하는 등의 문제점을 포함한 여러 한계를 지님을 확인하였다. 이를 바탕으로, 우리는 이러한 문제를 해결하기 위해 게이지(Gauge) 기반의 새로운 시각적 리셋 UI를 제안하고, 이를 기존 UI(Direction, End Point, Arrow Alignment)와 비교하여 실험 1을 통해 효과를 검증하였다. 또한, 리셋 인식률과 수행 효율을 높이기 위해 시청각을 결합한 멀티모달 리셋 UI를 제안하고, 실험 2를 통해 그 성능을 평가하였다. 우리는 본 연구는 물리적으로 가상공간을 보행하는 사용자들을 위한 새로운 멀티모달 리셋 UI 패러다임을 제시한다.

1. 서론

가상현실을 직접 보행하며 체험하는 방식은 높은 몰입감을 제공하고 멀미 발생 위험을 줄이는 데 효과적이다. 그러나 현실 공간과 가상 공간 간의 불일치로 인해 현실 공간의 장애물이나 경계면과 충돌할 위험이 존재한다. RDW 기술은 사용자의 HMD 화면을 미세하게 조정함으로써, 제한된 현실 공간 내에서도 보다 넓은 가상을 탐색하고 체험할 수 있도록 지원한다 [1]. 하지만

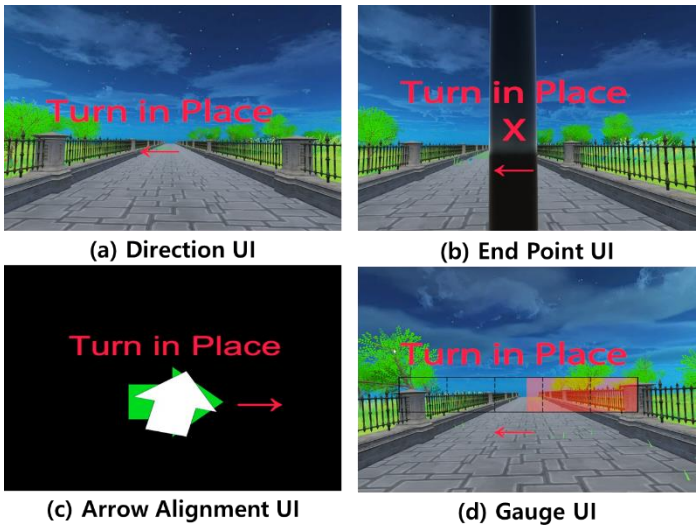
* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract)으로서, 원본 논문은 IEEE VR 2025에 발표 및 IEEE Transactions on Visualization and Computer Graphics 31(5). 2025에 게재되었음.
* 이 연구는 정부 (과학기술정보통신부)의 재원으로 한국연구재단 (No. RS-2024-00348094) 및 한국전파진흥협회 (No. RNIX20230200)의 지원으로 수행되었음.

그림 1: 시각, 청각, 시청각을 자극하는 리셋 UI

미세한 조정만으로는 충돌을 완전히 방지할 수 없는 경우, 사용자의 움직임을 제어하기 위해 HMD 화면에 UI를 표시하거나 음향 효과를 사용하는 방식이 필요하며, 이를 리셋(reset)이라고 한다 [2]. 최적의 리셋 방향을 결정하기 위한 resitter 연구와 함께 정확한 리셋 방향으로 사용자가 리셋을 행할 수 있도록 다양한 시각적 리셋 UI가 제안되어 왔으며, 어떤 방식이 VR 경험에 가장 적합한지는 충분히 검증되지 않았다. 이에 본 연구에서는 RDW 전문가들을 대상으로 인터뷰를 실시하여 기존 시각적 리셋 UI의 문제점을 파악하였으며, 이러한 문제를 해결하기 위해 새로운 시각 및 청각 기반 리셋 UI를 설계하고 사용자 실험을 통해 그 효과를 평가하였다([그림 1]).

2. 사전 연구

우리는 새로운 리셋 UI를 설계하기에 앞서, 기존 리셋 UI의 현황과 문제점을 조사하였다. 이를 위해 리셋 기술이 적용된 사용자 실험을 수행한 경험이 있는 4명의 전문가(모두 남성, 연령: 25~35세, RDW 연구 경력: 2년 이상)를 대상으로 인터뷰를 실시하였다. 인터뷰 결과, 사용자가 시스템이 제안한 최적의 리셋 방향으로 정확히 회전하지 못하는 경우가 빈번하다는 사실을 확인하였다. 또한, 사용자들이 눈앞에 표시된 UI를 즉시 인식하지 못한 채 보행을 지속하거나, 회전해야 하는 방향을 정확히 파악하지 못하는 문제도 드러났다. 마지막으로, 일부 사용자는 리셋 과정이 복잡하게 느껴져 전체 흐름을 이해하는 데 시간이 오래 걸렸다고 응답하였다. 우리는 이러한 인터뷰를 통해 밝혀진 문제점들을 해결하기 위해 새로운 리셋 UI를 제안한다.



(a) Direction UI (b) End Point UI

(c) Arrow Alignment UI (d) Gauge UI

그림 2: 실험 1의 각 UI가 적용되었을 때 사용자 화면

3. 연구 1

우리는 먼저 네 가지 시각적 리셋 UI를 비교 분석하였다([그림 2]). 리셋 UI는 사용자가 물리적 공간의 경계에 도달하여 리셋이 필요할 때, 사용자의 시야 앞에 생성된다. Direction UI는 사용자가 걸음을 멈추고 제자리에서 회전하도록 안내하는 문구 “Turn in Place”와, 회전 방향을 나타내는 화살표 기호로 구성된다. End Point UI는 Direction UI를 확장한 형태로, 사용자가 정확한 방향으로 회전할 수 있도록 종료 지점을 나타내는 실린더와 ‘X’ 기호를 함께 제시한다. Arrow Alignment UI는 사용자의 화면을 일시적으로 암전시키고, 시야 내에 흰색과 녹색 화살표를 통해 현재 사용자의 방향과 목표 방향을 각각 나타낸다. Gauge UI는 본 연구에서 새롭게 제안하는 방식으로, 리셋 진행 상황을 게이지 형태로 시야에 표시한다. 사용자 실험을 통해 각 UI 간의 가상 공간 탐색 성능 및 사용자 경험의 차이를 분석하였다. 탐색 성능 평가는 실험 중 평균 리셋 간 거리, 평균 리셋 각도 오차, 평균 리셋 위치 오차, 전체 실험 시간을 기록하여 수행하였다. 사용자 경험은 다음의 설문지를 활용하여 측정하였다: 멀미(SSQ), 몰입감(E2I), 현존감(IPQ), 시스템 사용성(SUS), 인지 부하(NASA-TLX). 각 조건에 대한 실험 종료 후에는 인터뷰를 통해 주관적인 사용자 경험도 평가하였다.

총 24명의 참가자가 실험에 참여하였으며, 각 UI 조건별로 80m의 가상 공간을 보행하도록 하였다. 수집된 데이터를 통계적으로 분석한 결과, 리셋 UI가 탐색 성능 및 사용자 경험에 유의미한 영향을 미친다는 사실을 확인하였다. 사후 분석 및 인터뷰를 통해 각 UI의 장단점을 파악했으며, Gauge UI가 기존 UI에 비해 성능이 개선되었음을 확인하였다. 특히 참가자 중 11(46%)명은 Gauge UI를 가장 선호한다고 응답하였다. 그러나 여전히 시각적 피로감이나 리셋 타이밍을 놓치는 등의 문제가 남아 있었기 때문에, 시각 이외의 감각을 활용한 새로운 UI 설계의 필요성이 제기되었다.

4. 연구 2

연구 2에서는 청각을 활용하여 리셋을 안내하는 새로운 UI를 제안하고, 연구 1에서 우수한 성능을 보인 Gauge UI와 비교하였다. 청각을 활용한 리셋 안내는 추가 장비 없이 HMD만으로 구현이 가능하다는 장점이 있다. 한편, 기존 연구에서는 VR 환경에서 시각과 청각 자극을 결합할 경우 사용자 경험이 향상된다는 결과가 보고된 바 있다. 이에 따라 본 연구에서는 시각과 청각을 결합한 멀티모달 리셋 UI를 실험에 포함하여 그 효과를 검증하고 사용자 경험 차이를 확인하였다([그림 1]).

청각 기반 리셋 UI는 리셋이 필요한 시점에 HMD의 스피커를 통해 간단한 비프음을 재생하도록 설계되었다. 사용자는 소리가 들리면 걸음을 멈추고, 지시된 방향으로 회전하게 된다. 이 UI는 회전 방향을 전달하기 위해 좌우 스테레오 사운드를 사용하며, 사용자가 돌아야 할 방향에 따라 왼쪽 또는 오른쪽 스피커를 통해 음향이 재생된다. 리셋 종료 지점에 가까워질수록 비프음의 템포가 점점 빨라져 리셋 상황을 직관적으로 전달하고, 리셋이 완료되면 완료음을 재생하여 피드백을 제공한다.

연구 1과 동일한 절차로 총 28명의 참가자를 대상으로 실험을 진행하였다. 실험 결과, 리셋을 안내하는 감각 유형에 따라 사용자 경험과 탐색 성능 모두에서 통계적으로 유의미한 차이가 나타났다. 또한 참가자 중 15(54%)명은 멀티모달 UI를 가장 선호하고 응당합였다. 특히, 시각과 청각을 결합한 멀티모달 UI는 각각의 장점을 효과적으로 통합하여, 사용자에게 가장 이상적인 리셋 경험을 제공함을 확인할 수 있었다.

5. 토의 및 결론

우리는 사용자에게 최적인 리셋 UI를 비교한 최초의 연구를 수행하였다. 전문가 인터뷰를 통해 기존 UI가 리셋 기술의 성능을 충분히 전달하지 못한다는 문제를 확인하였고, 이를 반영해 새로운 UI를 설계하였다. 본 연구에서는 시각 및 청각을 결합한 멀티모달 리셋 UI를 처음 제안하고, 기존 단일 감각 기반 UI보다 효과적으로 리셋을 유도함을 확인하였다. 향후에는 촉각 피드백 등 다양한 감각을 활용한 UI 확장을 통해, VR 환경 탐색과 사용자 경험의 질을 더욱 향상시킬 수 있을 것으로 기대된다.

참고문헌

[1] RAZZAQUE, Sharif. Redirected walking. The University of North Carolina at Chapel Hill, 2005.

[2] Williams, B., Narasimham, G., Rump, B., McNamara, T. P., Carr, T. H., Rieser, J., & Bodenheimer, B. (2007, July). Exploring large virtual environments with an HMD when physical space is limited. In Proceedings of the 4th symposium on Applied perception in graphics and visualization (pp. 41-48).

VR 아바타와 동일한 동작 수행에 대한 사용자 경험 분석 연구

이호정[0,1], 유상철[1], 고하영[1], 전윤석[1], 차승연[1], 이인권[1]
연세대학교 컴퓨터과학과[1]
{dearshawn, bestmark77, goha, seok220hun284, cha7367, iklee}@yonsei.ac.kr

User Experience Analysis Study on Performing the Identical Motion as VR Avatars

Ho Jung Lee[0,1], Sangcheol Yu[1], Hayoung Go[1], Yun Seok Jeon[1], Seung Yeon Cha[1], In-Kwon Lee[1]
Dept. of Computer Science, Yonsei Univeristy[1]

요약

본 연구는 VR 환경에서 사용자가 아바타를 직접 조작하여 가상 공간을 탐색할 때, 다양한 이동 동작 수행에 사용되는 조작 방식이 사용자 경험에 끼치는 영향을 분석하였다. 실험에 사용된 조작 방식은 Full-body tracking, Half-body tracking, Controller 기반의 세 가지이며, VR 환경 탐색을 위해 사용되는 대표적인 이동 동작들을 수행할 때 발생하는 피로도, 멀미, 몰입감, 사용성의 차이를 비교하였다. 연구는 외부 자극이 배제된 단순 환경과 실제 게임 시나리오 기반 환경에서 두 단계의 실험으로 구성되었으며, 실험 결과 조작 방식과 동작의 종류에 따라 사용자 경험에서 유의미한 차이가 발생함을 확인하였다. 이 결과는 향후 VR 콘텐츠 설계 시 사용자 친화적인 아바타 조작 체계를 수립하는 데 있어 중요한 기준이 될 것으로 기대된다.

1. 서론

VR 콘텐츠에서 아바타 조작 방식은 사용자 몰입도, 신체 피로도, 멀미 유발 등 경험 전반에 직결되는 중요한 요소이다. 특히 사용자의 전신 움직임을 반영하는 Full-body tracking, 상체만 추적하는 Half-body tracking, 손에 쥐고 있는 Controller의 버튼과 조이스틱을 이용하는 Controller 기반 방식은 가장 널리 활용되는 대표적인 조작 방식이다. 그러나 이러한 방식이 실제 VR 콘텐츠에서 다양한 이동 동작 수행 시 사용자에게 어떤 영향을 미치는지에 대해 종합적이고 체계적인 비교 연구는 충분히 이루어지지 않았다. 기존 연구에서는 손동작이나 음성, 키보드 등의 입력 방식을 활용한 아바타 제어 실험이 수행되었으나, 대부분 HMD 기반의 몰입형 환경이 아니거나, 전신 조작을 고려하지 않았으며, 실험에 사용된 동작의 수가 적어 실제 VR 콘텐츠에 적용하기에는 한계가 있었다 [1]. 또한, 일부 연구는 상체나 일부 신체 부위만을 사용하여 동물 아바타의 단순 동작을 반복하는 방식으로 조작 방식을 비교하였으며, 반복되는 복합 동작 속에서 발생하는 피로도나 멀미와 같은 생리적 반응을 충분히 측정하지 못했다 [2].

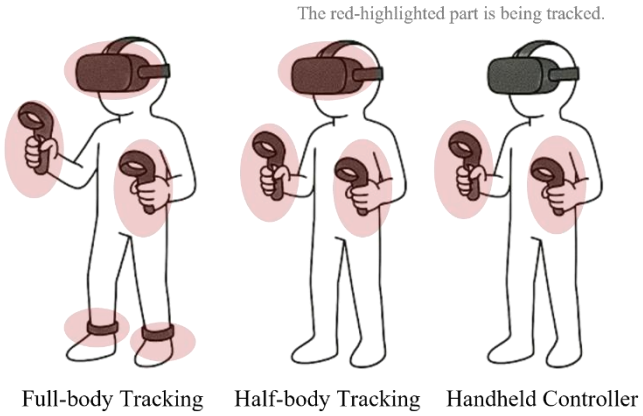


그림 1: 각 조작 방식의 개요

본 연구에서는 이러한 한계를 극복하고자 가상 환경 내에서 빈번하게 사용되는 대표 이동 동작들을 중심으로, 세 가지 조작 방식을 동일한 콘텐츠 환경 내에서 구현하고 비교하였다. 실험에는 걷기, 달리기, 앉기, 점프, 네발 기기의 대표적인 네가지 동작을 포함하였는데, 각 동작들은 서로 다른 신체의 부위를 사용하며 다양한 난이도를 갖는 동작이다. 참가자들은 동일한 VR 콘텐츠 내에서 각각의 조작 방식으로 모든 동작을 수행하며 실험에 참여하였다.

2. 조작 방식 설계

Full-body tracking 방식은 HMD와 컨트롤러, 그리고 발목에 부착된 트래커를 이용해 전신의 움직임을 1:1로 반영하는 방식으로 구현되었다. 이 방식은 실제 사용자의 동작을 그대로 재현하여 아바타가 움직이며, 네발 기기와 같은 고난도 자세도 직접 수행해야 하는 점에서 높은 신체적 개입을 요구한다. Half-body tracking은

* 구두 발표논문
* 본 논문은 요약논문(Extended Abstract)로 연구의 초기결과임.
* 이 연구는 정부 (과학기술정보통신부)의 재원으로 한국연구재단 (No. RS-2024-00348094) 및 한국전파진흥협회 (No. RNIX20230200)의 지원으로 수행되었음.

HMD와 컨트롤러만을 이용하여 상체 동작만 반영하며 구현되었다. Controller 방식은 모든 동작을 컨트롤러의 버튼과 조이스틱 입력만으로 수행하며, 실제 신체 움직임 없이 입력만으로 조작이 가능하도록 설계되었다. 세 방식 모두 Unity 엔진을 기반으로 동일한 VR 콘텐츠에서 통제된 조건 하에 구현되었으며, 참가자들은 실험 환경에 익숙해진 후 세 가지 조작 방식 동일한 동작을 수행하였다 ([그림1]).

3. 연구 방법

실험 1은 외부 시각 자극을 제거한 단순한 흰 배경 환경에서 진행되었다. 이 실험에서는 각 조작 방식에 따라 4가지 이동 동작(걷기, 앉기, 점프, 네발 기기)을 반복 수행하도록 하였으며, 각 동작 수행 직후 신체 부하(NASA-TLX)와 멀미 (FMS)를 평가하였다. 본 실험의 목적은 조작 방식과 이동 동작의 조합이 사용자 경험에 미치는 영향을 독립적으로 비교하는 데 있다.
실험 2는 실제 VR 콘텐츠 환경에서 각 조작 방식이 사용자 경험에 미치는 영향을 종합적으로 확인하기 위해 설계되었다. 실험 참가자들은 첫 번째 실험에서 각 이동 동작별로 가장 선호한 조작 방식을 선택하여 이를 조합한 맞춤형 방식(Custom)과, 기존의 3가지 단일 조작 방식을 비교하였다. 실험 참가자는 눈앞에 지속적으로 장애물이 등장하고, 점프, 앉기, 네발 기기 동작을 통해 장애물을 회피하며, 동시에 제자리에서 걷는 동작으로 아바타를 전진시키는 콘텐츠를 체험한다. 체험 종료 후에는 신체 부하(NASA-TLX), 멀미(FMS), 몰입감(IPQ), 사용성(SUS) 설문지를 바탕으로 경험 데이터를 수집하였다. 본 연구는 7명의 실험 참가자를 대상으로 진행한 파일럿 테스트 결과이다.

4. 실험 및 결과

4.1. 실험 1

수집된 표본수가 적기에 비모수 검정인 Kruskal-Wallis 테스트를 통하여 유의미성을 확인하였다. 피로도는 조작 방식 간에 유의미한 차이 ($p < .001$), 멀미는 이동 동작 간에 유의미한 차이 ($p < .05$)를 확인하였다. 사후 검정을 통하여 Full-body 방식은 다른 동작들보다 유의미하게 높은 피로도를 보였다. 또한, 멀미 측면에서 점프와 네발기기는 유의미한 차이가 나타났다. 이동 동작과 조작 방식은 멀미 및 피로도에 유의미한 영향을 끼치는 경향을 확인하였고 이는 충분한 표본을 모집하여 실험을 진행할 경우 더욱 두드러질 것으로 기대된다.

4.2. 실험 2

실험 2에서는 게임 환경 내에서도 피로도에서 유의미한 차이가 관찰되었으며($p < .001$), Controller 방식이 Full-body 방식보다 유의미하게 낮은 피로도를 보였다 ($p < .001$). 멀미, 몰입감, 사용성 항목에서는 통계적으로

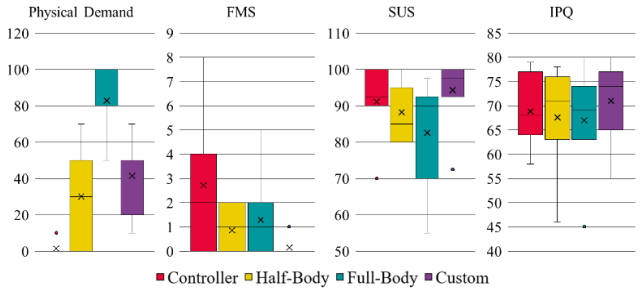


그림 2: 실험 2의 각 조작 방식에 따른 결과

유의한 차이는 나타나지 않았으나, 참가자들의 주관적 체감에서는 방식 간 차이가 보고되었으며, 샘플 수가 확대될 경우 유의미한 차이가 나타날 가능성이 있는 것으로 해석되었다 ([그림2]). 특히 그림2에서 Full-body tracking이 Controller보다 몰입감이 높은 양상은 이전 연구 [2]와 상반된 결과인데, 추가 실험에서 동작에 따른 새로운 결과 해석이 도출될 가능성이 있다. 맞춤형 방식의 경우, 전체적으로 피로도는 낮고 몰입감은 유지되는 긍정적인 반응을 보여 동작별 최적 방식 조합의 가능성을 시사하였다.

5. 토의 및 결론

조작 방식은 동작의 종류에 따라 사용자 경험에 유의미한 영향을 미치며, 특정 방식이 항상 우월한 것이 아니라 상황에 따라 조합할 필요가 있다. Full-body 방식은 이전 연구에서 몰입감이 높다는 결과를 보였으나, 반복 사용 시 높은 피로도와 멀미 유발 가능성이 존재하고 상대적으로 몰입감이 낮게 측정되었다. 또한 Controller 방식은 물리적 부담이 적고 반복 동작에 효과적이며, Half-body 방식은 그 중간 지점에서 몰입감과 부담을 균형 있게 제공한다. Tracking 장치 수를 줄임에 따라 발생하는 사용자 경험과 멀미, 피로도와 같은 생리적 반응의 상충 관계 속에서 사용자의 맞춤화된 조작 방식 선택은 각 요소측면에서의 단점을 해소할 수 있을 것으로 기대된다. 이러한 결과는 VR 콘텐츠 설계 시 사용자 친화적인 조작 체계를 구축하는 데 실질적인 기준이 될 수 있으며, 특히 동작의 종류와 빈도에 따라 조작 방식을 유동적으로 설계하는 방식이 효과적일 수 있음을 시사한다.

참고문헌

[1] JunSeo Park, Hanseob Kim, and Gerard Jounghyun Kim. 2024. A Study of 3D Character Control Methods_Keyboard, Speech, Hand Gesture, and Mixed Interfaces. In SIGGRAPH Asia 2024 Posters (SA '24). Association for Computing Machinery, New York, NY, USA, Article 1, 1–2.
[2] Andrey Krekhov, Sebastian Cmentowski, and Jens Krüger. 2018. VR Animals: Surreal Body Ownership in Virtual Reality Games. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (CHI PLAY '18 Extended Abstracts). Association for Computing Machinery, New York, NY, USA, 503–511.

다감각 자극이 가상환경 내 포털 인지에 미치는 영향

박시연, 조인호[0], 김선정
한림대학교 융합소프트웨어학과
{M23051, M25052, sunkim}@hallym.ac.kr

The Effect of Multisensory Stimuli on Portal Recognition in Virtual Reality

Siyeon Bak, Inho Jo[0], Sun-Jeong Kim
Department of Convergence Software, Hallym University

요약

본 연구는 가상현실 환경에서 감각 자극의 조합이 포털 너머 공간 인지에 미치는 영향을 분석하였다. 실험 결과, 감각 조건 간 응시 시간과 정답률에서 통계적으로 유의한 차이는 없었으나, 다감각 조건에서 빠른 반응과 높은 주관적 유용성 평가가 나타났다. 사후 인터뷰에서는 참가자들이 청각과 후각 자극을 전략적으로 활용하여 공간을 추론하는 경향이 확인되었다. 감각 자극은 단순한 보조 수단을 넘어 인지 전략에 기여할 수 있음을 시사한다.

1. 서론

인간은 오감을 통해, 주변 환경을 인지할 수 있다. 시각은 형태, 공간 그리고 청각은 거리와 방향성 정보, 촉각은 물체의 표면 질감을 알 수 있게 해주고 후각과 미각은 정서적 반응 및 기억 회상에 직접적으로 작용한다, 기존의 가상현실(Virtual Reality, VR) 환경은 주로 시각에 의존해왔으나, 최근에는 몰입감과 현실감을 증진시키기 위한 다감각 자극 활용에 대한 연구[1]가 활발히 이루어지고 있다. 특히 사용자가 제한된 시각 정보만으로 공간을 인지해야 하는 맥락에서는, 청각과 후각과 같은 비시각적 감각 정보가 인지 전략에 영향을 줄 수 있다. 본 연구에서는 포탈(Portal) 기반 VR 환경에서 다감각 자극이 사용자 공간 인지에 미치는 영향을 분석해보고자 한다.

2. 구현 및 설계

2.1. 구현

본 연구에서는 후각 자극을 제공하기 위해, Meta Quest 3 HMD에 부착 가능한 자체 개발형 후각 장치[2]를 활용하였다. 해당 장치는 총 네 가지 향을 분사할 수 있으며, 기기 측면에 장착되어 무선으로 제어된다.

가상환경은 Unity 6 엔진(버전 6000.0.32f1)을 기반으로 구현하였으며, 참가자는 정면에 위치한 포털을 마주하게 된다. 각 포털은 숲 또는 도시 중 하나의 장소로 연결되며, 해당 환경에는 서로 다른 수준의 배경음을 포함하고 있다. 또한, 또한 시각적 복잡성에 따라 사용자의 공간 인지에 미치는 영향을 보기 위해, 각 장소 내에 반복적으로 움직이는 물체의 크기, 위치 등을 달리해서 시각적 방해 요소를 조성하였다.

2.2. 실험 설계

본 실험은 감각 조건에 따라 참가자가 포털 너머에서 지나가는 물체를 인지하는 능력에 차이가 있는지를 검증하기 위해 설계되었다. 자극 자산은 감각 자극 유형에 따라 구성되었으며, 후각적 자극은 향을 기반으로 한 물체인 꽃, 오렌지, 피자, 사과의 향으로, 청각적 자극은 소리를 수반하는 물체인 트럭, 사람, 드론, 강아지의 소리로 설정하였다.

실험은 두 가지 공간(숲, 도시)(그림 1)을 배경으로 진행되었으며, 각 환경은 상이한 시각적·청각적 방해 요소를 포함한다. 숲은 상대적으로 낮은 환경 소음(바람, 새소리 등)과 움직이는 시각 요소(모닥불, 사슴)로 구성되었고, 도시는 높은 소음 환경(사람 소리, 경적, 엔진음 등)과 움직이는 시각 요소(헬리콥터, 앰뷸런스 등)로 구성되었다.

감각 조건은 다음의 네 가지로 구성되었다:
(1) 시각(V), (2) 시각 + 청각(VA), (3) 시각 + 후각(VO), (4) 시각 + 청각 + 후각(VAO).
청각 자극은 청각적 물체에 해당하는 소리를, 후각 자극은 후각적 물체에 해당하는 향을 동기화하여 제공하였다.
모든 참가자는 각 감각 조건을 총 16회씩 경험하였으며, 조건 순서는 라틴 스퀘어(Latin Square Design, LSD)로 무작위 순서로 제공되었다. 매 회마다 참가자는 제공된 감각 정보를 바탕으로 Portal 건너편에서 지나간 물체가 무엇인지 맞추도록 하였다. 선택은 VR 컨트롤러를 통해 이뤄졌으며, 응답시간(Portal 접근시점부터 응답 제출까지의 시간), 정답 여부, 자신감 평점(7점 척도)를 기록하였다.

그림 1. 가상환경 (좌: 숲, 우: 도시)

* 구두 발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 국제학술대회 발표심사 중임.
* 본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아서 수행된 연구임(No. RS-2023-00254695).

3. 실험

실험에는 총 16명의 참가자(남성 9명, 여성 7명)가 참여하였으며, 평균 연령은 25.75세(표준편차 = 2.67)였다. 실험 절차는 다음과 같이 구성되었다. 우선, 참가자는 연구 참여에 대한 동의서를 작성한 후, 인구통계학적 정보를 포함한 사전 설문지를 작성하였다. 이후 실험에 사용될 네 가지 향을 직접 시향하여 후각 자극에 대한 사전 경험을 제공하였다.

그림 2. 가상환경 내 포털

참가자는 의자에 착석한 상태에서 VR 디바이스와 컨트롤러를 착용하였으며, 본 실험에 앞서 UI 선택 방식과 조작법에 대한 간단한 훈련을 진행하였다. 본 실험은 각 감각 조건(총 4가지)별로 두 장소(도시, 숲)에서 각각 16회씩, 총 32회에 걸쳐 수행되었다(그림 2). 사전에 미리, 지나가는 물체는 하나, 혹은 여러 개 물체가 지나가며 지나간 물체가 무엇인지 UI 버튼에서 전부 선택하도록 지시하였다. 참가자는 Portal을 응시 후, 지나간 물체를 확인하여 물체와 유사한 아이콘을 선택하였다. 자극은 시각 외 청각·후각 조건에 따라 조건별로 달리 제시되었으며, 반응 시간, 정답 여부, 자신감 평정(7점 척도)을 기록하였다. 각 감각 조건이 종료된 후에는 SSQ(Simulator Sickness Questionnaire), NASA-tlx(NASA Task Load Index) 설문지로 SSQ를 작성하였으며, 조건 간에는 약 3분간의 휴식 시간이 주어졌다. 모든 실험이 종료된 뒤, 참가자는 사후 설문에 응답하였다.

4. 결과

감각 조건에 따른 포털 응시 시간과 정답률을 분석한 결과(표1), 다감각 조건에서 평균 응시 시간이 가장 짧았고, 후각 조건에서는 상대적으로 길게 나타났다. 그러나 도시 환경($F(3, 60) = 0.97$, $p = .413$)과 숲 환경($F(3, 60) = 0.11$, $p = .957$) 모두에서 감각 조건 간 차이는 통계적으로 유의하지 않았다. 정답률 또한 전반적으로 높은 수준을 유지하였으나, 도시($F = 0.20$, $p = .894$) 및 숲($F = 1.84$, $p = .150$) 조건 모두에서 유의한 차이는 확인되지 않았다.

표 1. 정답률과 응답시간

	V	VO	VA	VAO
정답률(%)	85.819	89.495	87.957	90.310
응답시간(초)	11.310	11.848	10.981	10.693

한편, 정량적 지표에서는 나타나지 않았던 감각 자극의 효과는 주관적 평가와 사후 인터뷰를 통해 명확히 드러났다. 참가자들은 감각 자극의 조합에 따라 서로 다른 인지 전략을 사용하였으며, 감각 정보가 풍부할수록 판단에 도움이 되었다고 응답하였다. 특히 주관적 유용성 평가에서는 다감각 조건이 평균 5.88점으로 가장 높은 점수를 받았고, 선호도 순위에서도 다감각 조건이 1순위로 가장 많이 선택되었다(그림 3).

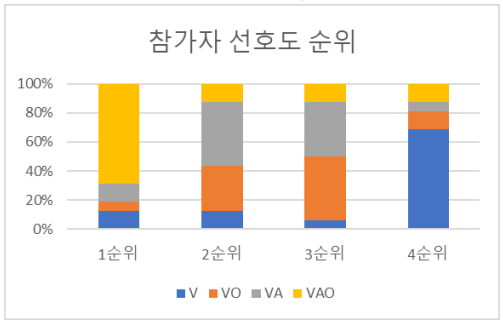


그림 3. 참가자 선호도 순위 그래프

사후 인터뷰에서는 청각 자극이 방향성을 예측하는 단서로, 후각 자극이 판단에 대한 확신을 강화하는 요소로 작용하였다고 응답하였다. 다수의 참가자들은 "소리로 예측하고, 시각으로 확인한 뒤, 향기로 확신했다"고 진술하며, 감각 간 정보를 전략적으로 통합하는 방식을 사용한 것으로 나타났다. 이는 감각 자극이 단순한 보조 정보가 아닌, 실제 인지 과정에 깊이 관여할 수 있음을 보여주며, 감각 통합의 효과는 정량적 분석만으로는 충분히 설명되지 않을 수 있음을 시사한다.

5. 결론 및 논의

본 연구는 가상현실 환경에서 감각 자극의 조합이 포털 너머 공간 인지에 미치는 영향을 확인하고자 하였다. 실험 결과, 감각 조건 간 응시 시간 및 정답률에서 통계적으로 유의한 차이는 나타나지 않았으나, 다감각 조건에서 전반적으로 빠른 반응과 높은 주관적 유용성 평가가 확인되었다. 사후 인터뷰에서는 참가자들이 시각 외 감각 자극을 전략적으로 활용하여 공간을 예측하는 경향이 나타났으며, 특히 청각과 후각이 주의를 유도하거나 확신을 제공하는 단서로 작용함을 알 수 있었다.

이러한 결과는 감각 자극이 단순한 보조 정보를 넘어, 인지 전략에 실질적으로 기여할 수 있음을 시사한다. 다만, 여전히 시각 정보에 크게 의존하는 경향이 관찰되었으며, 이는 결과 해석에 중요한 변수로 작용할 수 있다.

따라서 향후 연구에서는 시각 정보를 더욱 제한한 조건에서 청각 및 후각 자극이 공간 인지에 정량적으로도 유의미한 기여를 할 수 있는지를 추가로 검증할 필요가 있다.

참고문헌

[1] Yildirim, M., Globa, A., Gocer, O., & Brambilla, A. (2025). Digital Smell Technologies for the Built Environment: Evaluating Human Responses to Multisensory Stimuli in Immersive Virtual Reality. *Building and Environment*, 112608.

[2] Inho Jo, Siyeon Bak, & Sun-Jeong Kim (2024-07-09). Improvement of a wearable olfactory device to enhance VR user experience. 한국컴퓨터그래픽스학회 학술대회.

논문 발표

HCI/시각화/시스템

핵융합 마그네틱 아일랜드 탐지를 위한 시뮬레이션 데이터 생성 및 가시화*

김준호[0,1], 나민태[1], 윤세진[1], 윤의성[2], 김종현[3], 김선정[1]
[1]한림대학교, [2]울산과학기술원, [3]인하대학교
{M25050, M24052, M25051}@hallym.ac.kr, esyoon@unist.ac.kr, jonghyunkim@inha.ac.kr, sunkim@hallym.ac.kr

Generation and Visualization of the Simulation Data for Detecting the Magnetic Islands in Nuclear Fusion

Junho Kim[0], Mintae Na[1], Sejin Yun[1], Eisung Yoon[2], Jong-Hyun Kim[3], Sun-Jeong Kim[1]
[1]Hallym University, [2]Ulsan National Institute of Science and Technology, [3]Inha University

요약

본 연구는 핵융합 분야에서 흔히 쓰이는 물리 기반 PDE(편미분 방정식) 솔버 대신 함수 블렌딩 방식을 이용해 전자온도 평탄화 현상을 재현하고, 이를 바탕으로 마그네틱 아일랜드를 시뮬레이션 하였다. 또한, 상용 게임 엔진에서 실시간으로 가시화하기 위해 렌더 타겟과 컴퓨트 쉐이더를 사용하는 통합 파이프라인을 제안한다.

1. 서론

핵융합 실험 중 발생하는 마그네틱 아일랜드는 플라즈마를 가두는 자기장 중 일부가 섬처럼 나타나는 현상으로, 플라즈마를 가두던 닫힌 자기장 면들을 서로 연결해 플라즈마의 공간적 혼합을 활성화하여 핵융합 장치의 가둠 성능을 저해하는 주요 요인으로 꼽힌다[1].

그러나 현재 간접적으로 마그네틱 아일랜드 진단을 위해 사용되는 전자 온도 이미지 진단 장치(Electron cyclotron emission imaging)는 공간 해상도와 설치 위치에 제약이 있고, 마그네틱 아일랜드의 크기와 동적 특성 등 다양한 한계점으로 인해 실험 중 실시간 탐지 및 정확한 형상 파악에는 상당한 어려움이 따른다. 따라서 시뮬레이션을 통해 마그네틱 아일랜드 발생 과정을 모사하고, 이를 기반으로 생성된 실험 데이터를 시각화함으로써 탐지 정확도를 높일 필요가 있다[2].

본 연구는 이러한 문제를 해결하기 위해 실험 데이터 생성과 언리얼 엔진 기반 시각화를 결합하여 마그네틱 아일랜드를 효과적으로 탐지·분석하는 기법을 개발하는 것을 목표로 한다.

* 구두 발표논문
* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아서 수행된 연구임(No. RS-2023-00254695).
* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155915, 인공지능융합혁신인재양성(인하대학교))

2. 마그네틱 아일랜드 실험 데이터 생성

본 연구에서는 전자 온도 T_e가 단일 (m,n) 마그네틱 아일랜드가 생겼을 때 어떻게 평탄화 되는지를 수치적으로 모사하기 위해, 완전한 MHD/gyrokinetic 방정식이 아니라 간단한 함수 블렌딩을 사용한다.

먼저 아일랜드가 없을 때의 배경 프로파일 $T_{e,bg}(r)$은

$$T_{e,bg}(r) = \exp\left[-\kappa_T \tanh\left(\frac{r - x_0}{w}\right)\right], (1)$$

형태로 정의하며, 여기서 κ_T는 프로파일이 얼마나 급격하게 안쪽에서 바깥쪽으로 변화할지를 조절하는 무차원 파라미터, x_0는 T_e가 가장 가파르게 떨어지기 시작하는 중심의 반경위치, w는 반경 방향으로 tanh 전이 구간(Transition region)의 반너비(Half-width)이다.

다음으로, 마그네틱 아일랜드 내부에서는 평행 수송(Parallel transport)에 의해 T_e가 배경 프로파일 값 $T_{e,bg}(r_0^{(mi)})$로 완전히 평탄화된다고 보고, 이를 매끄럽게 연결하기 위해,

$$\kappa_{blend}(r) = \exp\left[-\left(\frac{r - r_0^{(mi)}}{w_{2D}(\theta,\varphi)}\right)^8\right], (2)$$

와 같이 가중치 함수를 정의한다.

실제 토카막 단면에서는 마그네틱 아일랜드의 반너비가 플로이달 각 θ와 토로이달 각 φ에 따라 바뀌므로

$$w_{2D}(\theta,\varphi) = w_0^{(mi)}\left|\cos\left(\frac{m\theta}{2} - n\varphi\right)\right|, (3)$$

이를 모사하기 위해, $(x,y) \mapsto (r,\theta)$ 치환 후 $r = \sqrt{x^2+y^2}$와 $\theta = \arctan 2(y,x)$를 계산하여 $W_{2D}(\theta,\varphi)$를 얻은 뒤, 최종적으로

$$T_e(r,\theta,\varphi) = \kappa_{blend}(r)T_{e,bg}(r_0^{(mi)}) + [1-\kappa_{blend}(r)]T_{e,bg}(r), (4)$$

와 같이 2차원 전자온도를 설정한다.

이처럼 실험적/이론적 지배 방정식을 풀지 않고도, 국소적으로 평탄화된 O-point 부분과 이를 감싸는 배경 프로파일을 매끄럽게 연결함으로써, 전형적인 (m,n) 마그네틱 아일랜드에 의한 온도 평탄화(Electron

temperature flattening)를 빠르게 생성할 수 있다.

본 연구에서는 이 방법을 사용하여 다양한 φ 면과 m, n 값을 변화시키면서 파라미터 스캔 및 데이터 분석 알고리즘 검증을 수행하였다.

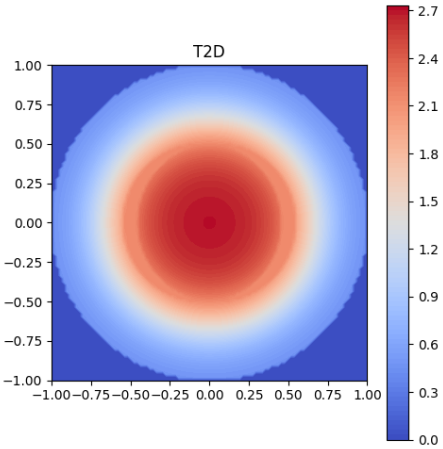


그림 1: 마그네틱 아일랜드 데이터의 파이썬 시각화

3. 언리얼 엔진 시각화

3.1. 온도 텍스처 생성 및 렌더 타겟 복사

앞서 생성된 데이터를 언리얼 엔진 환경에서 실시간으로 시각화하기 위해, 전용 액터를 구현하였다. 해당 액터는 실행 시 먼저 플라즈마 단면 온도 값을 텍스트 파일로부터 읽어 들이고, 이를 GPU 가속을 활용하여 원하는 해상도로 업스케일 및 Viridis 컬러맵을 적용한 뒤 UTexture2D 텍스처로 생성한다. 생성된 컬러 텍스처는 곧바로 UTextureRenderTarget2D로 복사되며, 이 렌더 타겟을 통해 UI 위젯 또는 3D 메시에 결과를 띄울 수 있다.

3.2. GPU 컴퓨트 쉐이더 기반 엣지 검출

이후, 엣지(edge) 검출을 위해 컴퓨트 쉐이더 파이프라인을 거친다. 해당 파이프라인은 Grayscale → Horizontal Sobel → Vertical Sobel → Gradient 합성(Combine) → Threshold(임계값 처리) → Outline(윤곽선 생성)의 순서로 실행되며, 모든 단계가 GPU 계산으로 병렬 처리된다. 이에 따라 프레임 레이트 저하 없이 실시간으로 경계선을 검출할 수 있다.

3.3. 시각화 모드

최종적으로, 엣지 검출 결과는 그림 2와 같이 "Overlay" 또는 "Blending" 두 가지 모드로 사용자에게 제공된다. Overlay 모드는 경계선 픽셀을 원본 온도 텍스처 위에 반투명 혹은 불투명으로 덧씌워 보여주는 방식이며, Blending 모드는 경계 픽셀을 원본 컬러맵과 일정 비율로 섞어 자연스럽게 강조하는 방식이다. 또한, 그림 3과 같이 정지 이미지뿐만 아니라 프레임마다 컴퓨트 쉐이더를 반복 실행하여 동영상처럼 연속된 시각화를 지원하므로, 실시간 데이터 분석에서도 유용하다.

이러한 모듈화 설계를 통해, 연구자는 파라미터(예: 엣지 임계값 등)를 블루프린트 또는 UI 슬라이더로 동적으로 조정할 수 있으며, 모든 연산(업스케일링, 컬러맵, 엣지 검출)은 GPU 컴퓨트 쉐이더를 통해 최적화되어 있다. 따라서 플라즈마 단면의 전자온도 분포를 실시간으로 탐색·분석하고, 즉시 화면에 확인할 수 있는 환경을 구현하였다.

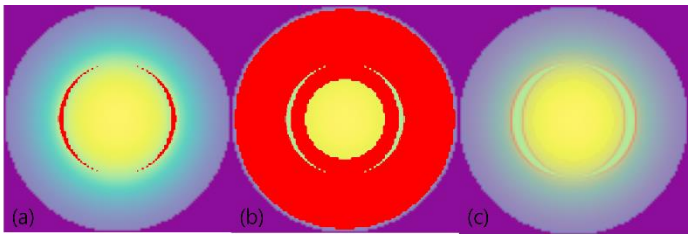


그림 2: 언리얼 시각화 (이미지 탐지), (a) Overlay 모드, (b)Overlay 모드(반전), (c) Blend 모드

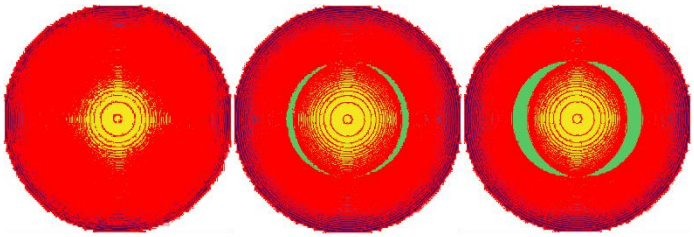


그림 3: 언리얼 시각화 (동영상 탐지), Overlay모드, 왼쪽부터 각각 $t = 0, t = 50, t = 100$ (초)

4. 결론 및 향후 연구

본 연구에서는 핵융합 플라즈마 내 마그네틱 아일랜드 실험 데이터를 생성하고, 이를 언리얼 엔진을 사용해 시각화하여 마그네틱 아일랜드를 효과적으로 탐지·분석하는 기법을 제안하였다. 그러나 다음과 같은 한계점이 남아 있다. 첫째, 실험 데이터가 마그네틱 아일랜드의 동적 변화(예: 회전, 충돌 등)를 충분히 반영하지 못하였으며, 둘째, 렌더링 시간 및 시스템 오버헤드에 대한 정량적 평가가 이루어지지 않았다.

향후 연구에서는 저해상도 데이터의 응용 가능성을 검증하고, AI 기반 학습을 도입하여 탐지 정확도를 향상시키는 한편, 노이즈 제거 기법을 추가하여 시각화 품질을 개선할 예정이다.

참고문헌

[1] Kwon, J.-M., Ku, S., Choi, M. J., Chang, C. S., Hager, R., Yoon, E. S., Lee, H. H., and Kim, H. S., Gyrokinetic simulation study of magnetic island effects on neoclassical physics and micro-instabilities in a realistic KSTAR plasma, *Physics of Plasmas*, 25(5): 052506, 2018.

[2] Tae, W., Yoon, E. S., Hur, M. S., Choi, G. J., Kwon, J.-M., and Choi, M. J., Unveiling non-flat profiles within magnetic islands in tokamaks, *Physics of Plasmas*, 31(2): 020702, 2024

단일 단계 B-rep 생성 확산 모델*

이민기[01], 장동수[2], 클레망 잠봉[3], 김영민[12]
[1]서울대학교 전기정보공학부
[2]서울대학교 뉴미디어통신 연구소
[3]매사추세츠 공과대학교 컴퓨터과학 및 인공지능 연구소
{mingi1019, 96lives, youngmin.kim}@snu.ac.kr, cjambon@mit.edu

BrepDiff: Single-stage B-rep Diffusion Model

Mingi Lee[01], Dongsu Zhang[2], Clément Jambon[3], Young Min Kim[12]
[1]Dept. of Electrical and Computer Engineering, Seoul National University
[2]Institute of New Media and Communications, Seoul National University
[3]Computer Science and Artificial Intelligence Laboratory(CSAIL), Massachusetts Institute of Technology

요약

B-rep(경계 표현)은 CAD 소프트웨어로 설계된 대부분의 디자인에서 널리 사용되는 3D 모델 표현 방식이다. 그러나 그 불규칙하고 희소한 기하, 위상 정보와 구조는 B-rep에 특화된 생성 모델 설계에 상당한 어려움을 제시한다. 이러한 문제를 해결하기 위해 단일 단계 B-rep 생성 확산 모델인 BrepDiff를 제안한다. 본 방법론은 면에서 구조화된 점 샘플로 구성된 마스킹된 UV 그리드 표현을 사용하며, 이를 확산 트랜스포머의 입력으로 활용한다. 또한 UV 그리드의 분포를 더 잘 포착하기 위해 비동기적 노이즈 스케줄을 도입하여 훈련 신호를 강화한다. 마스킹된 UV 그리드 표현의 명시성은 사용자가 위상적 유효성에 구애 받지 않고 표면 기하를 직관적으로 이해하고 자유롭게 설계할 수 있게 한다. BrepDiff는 형상 완성, 병합, 보간 등 복잡하고 다양한 기하 및 위상 조작을 가능하게 한다.

1. 서론

본 논문에서는 단일 단계 B-rep 생성을 위한 확산 모델인 BrepDiff를 제안한다. 마스킹된 UV 그리드 표현을 활용함으로써, 해당 방법론은 그림 2와 같이 모든 기하학적 구성 요소를 동시에 생성할 수 있어 다단계 파이프라인[1]의 필요성을 제거하고 더 풍부한 편집 기능을 가능하게 한다. 본 논문에서 제시하는 맞춤형 노이즈 스케줄링과 비동기적 2단계 노이즈 제거 전략은 마스킹된 UV 그리드 표현의 고유한 특성을 반영하여 생성형 확산 모델 훈련 효율성과 모델 정확도를 향상시킨다. BrepDiff는 고품질 B-rep 생성뿐만 아니라 광범위한 유연하고 강력한 편집 작업을 지원함을 입증한다.

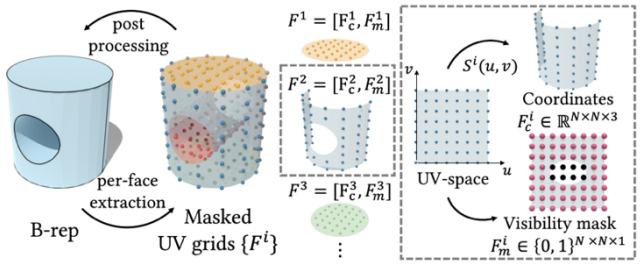


그림 1. 마스킹된 UV 그리드 표현의 예시.

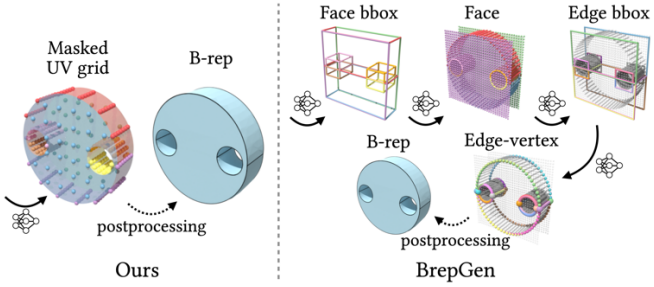


그림 2. BrepDiff의 전체 구조와 다단계구조인 BrepGen[1]의 전체 구조

2. 마스킹된 UV 그리드 표현

B-rep의 구조를 효과적으로 표현하기 위해 마스킹된 UV 그리드라는 표현 방식을 도입한다. 이 방법은 각 면의 기하학적 특성과 위상 정보를 동시에 포착하면서도, 딥러닝 기반 생성 모델과의 호환성을 유지하는 것이 특징이다.

하나의 B-rep에 대해, 각 i번째 면을 N×N(N=8)의 규칙적인 간격으로 샘플링한 이산화된 UV 그리드(F^i)로 표현한다(그림 1). 구체적으로, 각 면은 매개변수를 기준으로 한 UV도메인에서 함수 S^i를 통해 3차원 공간으로 대응되는 3차원 좌표 그리드(F_c^i)를 생성한다. 이 좌표와 함께, 각 그리드 포인트가 가시 표면의 일부인지 구멍이나 오목부에 가려진 것인지를 나타내는 이진 가시성 마스크(F_m^i)를 계산한다. 이 두 정보를 통합하여 마스킹된 UV 그리드는 면의 명시적 기하적 구조와 암

* 구두 발표논문
* 본 논문은 요약논문(Extended Abstract)으로서, 본 논문의 원 본 논문은 SIGGRAPH 2025에 발표 될 예정임.

시적 위상 정보를 동시에 포착한다.

신경망 관점에서 B-rep은 각 면당 하나의 토큰이 모인 집합으로 표현된다. 다양한 솔리드 모델 간에 면의 개수가 다르기 때문에, 토큰 집합을 고정 길이로 패딩하고, 트랜스포머 내 어텐션 마스크를 사용해 미사용 토큰을 무시한다. 이러한 접근법은 현대 신경망 아키텍처와의 호환성을 보장하며, 가시성 마스크를 통해 명시적 에지나 정점 생성 없이도 구멍과 같은 위상적 특징을 표현할 수 있다.

3. 확산 모델

생성 과정은 면 토큰 집합에 적용된 가우시안 확산 과정을 역으로 학습하는 확산 트랜스포머로 모델링한다. 이미지 확산 모델에서 흔히 사용되는 표준 노이즈 스케줄은 이미지보다 훨씬 적은 신호를 담고 있는 UV 그리드의 기하학적 신호를 빠르게 저하시켜 노이즈 제거를 어렵게 만든다. 이를 해결하기 위해, log SNR이 선형으로 감소하는 스케줄을 도입한다. 이 스케줄은 초기 단계에서 높은 SNR을 유지해 기하학적 세부 사항을 보존하며, 의미 있는 노이즈 제거 경로 학습을 용이하게 한다. 또한 좌표와 가시성 마스크의 학습 역학이 상이함을 관찰하여, 비동기적 2단계 노이즈 제거 전략을 채택한다. 초기 단계에서는 모델이 기하학적 정보에 집중하고, 후기 단계에서는 기하학과 가시성 마스크를 공동으로 예측한다. 이렇게 두 단계로 분리함에 따라 전체 형상 확립 후 미세 위상 세부 사항 해결을 가능케 하여 생성 안정성과 품질을 향상시킨다.

4. 후처리

생성된 마스킹된 UV 그리드 집합은 기하학 기반 후처리 파이프라인을 통해 실제 B-rep으로 변환한다. 먼저 가시 표면에 있는 점 샘플로부터 푸아송 표면 재구성[2]을 통해 점유 체적을 복원한다. UV 그리드는 경계 패딩을 통해 확장되며, 결과 메시는 자기 교차를 해결하고 독립된 표면 패치를 형성하도록 분할한다. 일반화된 감김 수(winding number)를 이용한 견고한 내부-외부 분할 기법[3]을 통해 점유 체적을 둘러싼 패치를 식별한 후, 이 패치들의 교차 계산을 통해 B-rep의 에지와 정점을 추출한다. 이 과정은 입력 UV 그리드와 결과 위상을 모두 반영하는 견고한 B-rep을 생성한다.

5. 결과

BrepDiff는 DeepCAD[4] 및 ABC[5] 데이터셋에서 학습한다. 무작위 생성 결과(그림 3)는 자유 곡면과 복합 위상 구조를 포함하는 유효한 B-rep 생성을 보여준다. 다단계 구조인 BrepGen은 솔리드 당 96.64초가 걸리지만, 단일 단계인 BrepDiff는 솔리드 당 13.55초로 훨씬 빠르다. 1-NNA 지표는 BrepGen 60.3, BrepDiff 60.67로 거의 차이가 없다.

또한, 명시적 토큰 기반 표현과 확산 모델의 유연성을 통해 사용자의 유연하고 다양한 편집 기능을 가능하게 한다. 본 모델의 편집 기능의 예시인 솔리드 자동완성, 병합, 보간은 두 가지 핵심 메커니즘으로 구현된다. 첫째, 형상 완성과 병합은 지정한 부분 UV 그리드의 토큰들에 노이즈를 주입한 후 노이즈 제거 과정에 계속 삽입해주며 미지정된 부분을 자연스럽게 생성하는 방식이다. 둘째, 형상 간 보간은 DDIM 역전을 통해 획득한 잠재 공간에서 최적 수송 기반 토큰 매칭을 수행한 후 보간한다. 보간 된 토큰을 다시 DDIM을 통해 샘플링하면 자연스럽게 보간 된 형상이 생성된다. 그림 4는 위 과정을 적용하여 실제로 편집을 수행한 결과를 보여준다.

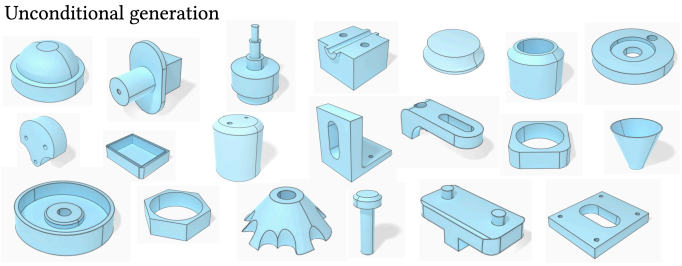


그림 3. 무작위 생성 예시

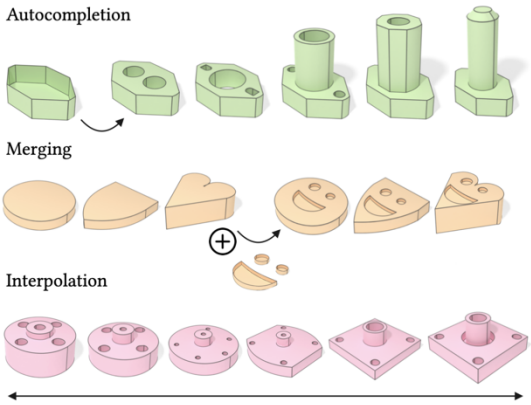


그림 4. 자동완성, 병합, 보간의 예시

References

[1] Xiang Xu, Joseph G. Lambourne, Pradeep Kumar Jayaraman, Zhengqing Wang, Karl D.D. Willis, Yasutaka Furukawa. "BrepGen: A B-rep Generative Diffusion Model with Structured Latent Geometry." *ACM Transactions on Graphics* (2024)

[2] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. "Poisson surface reconstruction." *In Proceedings of the Fourth Eurographics Symposium on Geometry Processing* (2006)

[3] Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. "Robust inside-outside segmentation using generalized winding numbers." *ACM Transactions on Graphics* (2013)

[4] Rundi Wu, Chang Xiao, Changxi Zheng. "DeepCAD: A Deep Generative Network for Computer-Aided Design Models." *In Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)*

[5] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, Daniele Panozzo. "ABC: A Big CAD Model Dataset For Geometric Deep Learning." *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019)

원격 상호작용 시 뇌 동기화 연구: Embodied 로봇 매개 상호작용을 중심으로*

강민규[0,a], 유재환[a], 정면걸[b], 김광욱[†,a]
[a]한양대학교 컴퓨터소프트웨어학과, [b]한양대학교 의류학과
{rkdalsrb3479, jaehwanyou0724}@gmail.com, {wjdausrjf, kenny}@hanyang.ac.kr

Research on Inter Brain Synchrony in Robot-Mediated Remote Social Interaction

Mingyu Kang[0,a], Jaehwan You[a], Myeongul Jung[b], Kwanguk(Kenny) Kim[†,a]
[a]Dept. of Computer Science, Hanyang University, [b]Dept. of Clothing & Textiles, Hanyang University

Abstract
본 연구에서는 Embodied Robot을 매개로 한 원격 상호작용 시의 사전 물리적 동기화가 사회적 상호작용과 뇌 동기화도에 주는 영향에 대해 연구하였다. 동기화와 비동기화 조건에서 원격 교육 과제를 수행한 결과, 매개 로봇과의 사전 물리적 동기화가 사회적 친밀감과 우측 전전두피질의 뇌 동기화도를 유의미하게 높임을 확인했다. 이는 휴머노이드를 매개로 한 원격 상호작용의 품질을 물리적 동기화가 높일 수 있음을 시사한다.

1. Introduction
원격 상호작용은 참여자들이 공간적 한계를 뛰어넘어 소통할 수 있게 하는 기술이다. 최근에는 매개 로봇을 도입하여 물리적 및 사회적 상호작용을 더욱 강화하는 방식도 연구되고 있다. 특히 원격지에 있는 참여자의 전신 움직임을 실시간으로 모방하는 휴머노이드인 체화된 로봇(embodied robot, ER)은 체화(embodiment), 공존감(copresence), 감정(valence) 등 측면에서 원격 사회적 상호작용의 품질을 높일 수 있음이 연구되었다[1].

상호작용의 품질을 평가할 수 있는 객관적·생리적 지표로는 뇌 동기화도(inter brain synchrony, IBS)가 있다. IBS란 뇌 활동의 동시성과 유사성을 나타내는 지표로, 특히 전전두피질(prefrontal cortex, PFC)에서의 IBS가 사회적 상호작용의 품질을 잘 반영하는 것으로 알려져 있다[2]. PFC의 IBS를 원격 상호작용에 대해 조사한 기존 연구들에 따르면 IBS는 개별/경쟁 작업보다는 협력 작업에서, 음성 상호작용보다는 화상 상호작용에서 높게 나타난다[3, 4]. 또한 원격 교육 상황을 다룬 연구에서는 교육에 앞서 선생과 학생이 AR 아바타를 통해 물리적으로 동기화(prior physical synchrony, prior PS)되면 IBS가 높아질 수 있음을 확인했다[5]. 이를 기초로 원격 상호작용에서도 매개 로봇과의 prior PS가 영향이 있을 것이라는 가설을 수립하였고, 본 연구에서는 ER을 매개로 하는 원격 교육을 설계해 prior PS 여부에 따른 사회적 친밀감과 IBS를 연구하고자 한다.

* 구두발표논문
* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00355411, 가상 아바타와 사람간의 Embodiment 형성을 활용한 응용 기술 개발 및 심리학적 영향 평가).
† Correspondence to K. Kim (Kenny) (kenny@hanyang.ac.kr)

2. Method
2.1. Participants
학생 역할을 맡을 피험자를 총 16명(12명 남성) 모집하였으며, 평균 나이는 25.81세(SD=4.62)였다.

2.2. Apparatus
PFC의 뇌 활동을 측정하기 위해 15개 채널로 구성된 fNIRS LITE (OBELAB, Korea)를 사용했다. ER로는 25의 자유도를 갖는 휴머노이드인 NAO v6 (Softbank, Japan)를 사용했고, Unity 엔진과 Motive (Optitrack, USA) 모션캡처 시스템을 통해 선행연구에서 제안한 실시간 동작 모방 시스템을 구현하였다[1].

2.3. Task
기존 연구[3]에서 제안한 원격 정보 전달 상호작용 과제를 사용하였다. 이 과제는 movement와 learning, testing을 순서대로 수행하는 구성으로, 실험자가 선생, 피험자가 학생 역할을 맡는다. 두 참여자의 공간을 분리하여 학생은 ER만 볼 수 있고, 선생은 ER의 내장 카메라를 통해 모니터로 학생을 볼 수 있도록 설계하였다.

Movement phase (6분)에서는 선생(ER)과 학생이 마주본 상태에서 신호음에 맞춰 허공에 가상의 원을 그리는 작업을 수행했다. 둘은 신호음(23/27 bpm)이 들릴 때마다 사전에 지정된 방향(시계/반시계)으로 원의 1/8씩 손을 이동시켰다. 동기화 조건(PS)에서는 신호음의 주기는 동일하게, 방향은 반대로 제시하여 거울상이 되도록 하였다. 비동기화 조건(physical asynchrony, PA)에서는 서로 다른 주기의 신호음과 동일한 방향을 제시하였다.

Learning phase (6분)에서는 선생의 낭독을 듣고 8개 단어를 암기하는 원격 교육을 수행했다. 선생과 학생은 각자 모니터를 통해 단어카드를 제시받았다. 학생에게는 단어와 발음만 표시되었고 선생에게는 단어, 발음에 더해 뜻과 예문이 함께 표시되었다. 선생은 카드마다 40초에 걸쳐 단어와 발음, 예문을 반복해서 낭독했고, ER을 통해 학생과 눈을 맞추며 비언어적으로 소통했다.

Testing phase에서는 learning phase에서 암기한 8개 단어에 대해서 2분간 단어 시험을 치렀다. 시험은 단어의 한국어 뜻을 쓰는 문항과, 예문의 빈칸에 들어갈 단어를 쓰는 문항으로 구성되었다 (총 16문제).

2.4. Procedure
피험자들은 사전설문을 작성한 후 fNIRS를 착용, PS와 PA 조건으로 task를 수행했다. 두 조건의 순서는 역균형화되었고 각 task가 종료된 후에는 설문을 작성했다.

2.5. Dependent Measures
(1) 사회적 친밀감
피험자들은 각 task가 종료된 직후, 기존 연구들에서 사용한 설문지를 통해 친밀감(rapport), 자신감, 공존감, 사회적 실재감(social presence)을 각각 7점, 7점, 5점, 5점 척도로 자기 평가하였다.

(2) Inter Brain Synchrony (IBS)
생체 노이즈를 제거하기 위해 learning phase에서 측정된 뇌 신호를 0.2Hz의 low pass filter와 0.02Hz의 high pass filter로 전처리하였고, 이를 바탕으로 채널별 wavelet transform coherence (WTC)를 계산하였다.

(3) Quiz Score
피험자들이 상호작용에 집중하였는지를 확인하기 위하여 각 단어에 대해 두 가지 문항을 모두 맞힌 경우에만 1점이 주어지는 방식으로 채점하였다(최대 8점).

3. Results
3.1. Social Closeness
PS와 PA 조건에 대한 t-test 결과 4종의 설문지 중 친밀감(p=.044)과 공존감(p=.039)은 PS 조건에서 유의미하게 높게 나타났다. 자신감(p=.296)과 사회적 실재감(p=.086)은 유의미한 차이를 보이지 않았다 (그림1).

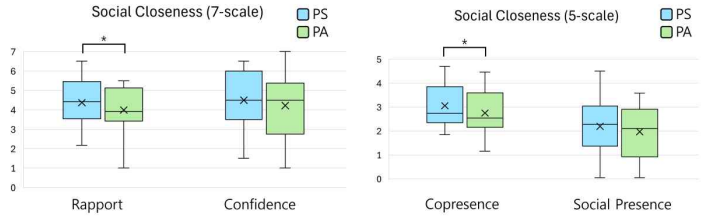


그림 1: 사회적 친밀감

3.2. Inter Brain Synchrony (IBS)
(1) Average IBS of 15 Channels
Average IBS에 대해 PS 조건(M=0.374, SD=0.033)에서가 PA(M=0.357, SD=0.037)에서보다 높았고, t-test 결과 p=.076으로 확인되었다 (그림 2).

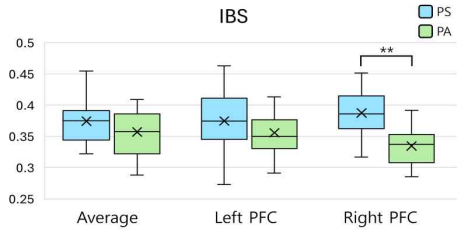


그림 2: Inter Brain Synchrony

(2) Left vs. Right IBS
PFC를 좌측(left PFC, lPFC)과 우측(right PFC, rPFC)으로 나누어 비교하였다. lPFC에서는 PS 조건(M=0.375, SD=0.048)에서가 PA(M=0.356, SD=0.040)보다 높되 유의미한 차이는 없었으나(p=.141), rPFC에서는 PS(M=0.383, SD=0.040)에서가 PA(M=0.347, SD=0.053)보다 유의미하게 높았다(p=.008).

3.3. Quiz Score
t-test를 실시한 결과, PS 조건(M=5.06, SD=3.15)과 PA 조건(M=5.19, SD=2.88) 간에서 유의미한 차이는 발견되지 않았다(p=.779).

4. Discussion
본 연구에서 사회적 친밀감의 결과는 매개 ER과의 prior PS가 사회적 상호작용의 품질을 높일 수 있음을 시사한다. 이는 모션캡처 기반의 motion planning을 하는 ER을 통해 원격 사회적 상호작용이 가능하다는 기존 연구 결과[1]와 결을 같이하며, 기존에 AR 아바타를 통해 확인되었던 prior PS의 효과[3]가 매개 휴머노이드를 통해서도 나타날 수 있음을 보여준다. IBS의 결과는 사회적 친밀감의 결과를 뒷받침한다. 특히 시각적 단어 인식과 음성 수신 및 해석, 감정 조절에 관여하는 것으로 알려진[6] rPFC에서의 유의미한 차이는 prior PS가 해당 뇌 활동을 보다 활성화했을 가능성을 암시한다. 또한 이것은 IBS가 원격 사회적 상호작용 평가의 척도로 활용될 가능성을 시사한다.

본 연구에서는 휴머노이드를 매개로 한 원격 사회적 상호작용의 품질을 사전 물리적 동기화가 높일 수 있음을 보였다. 후속 연구에서는 휴머노이드를 매개로 한 사회적 상호작용 평가 척도로서의 IBS에 대해 추가적인 검증이 이루어져야 할 것이다.

참고문헌
[1] Jung, M., Kim, J., Han, K., & Kim, K., Social telecommunication experience with full-body ownership humanoid robot, *International Journal of Social Robotics*, 14(9), 1951-1964, 2022.

[2] Czeszumski, A., Liang, S. H. Y., Dikker, S., König, P., Lee, C. P., Koole, S. L., & Kelsen, B., Cooperative behavior evokes interbrain synchrony in the prefrontal and temporoparietal cortex: a systematic review and meta-analysis of fNIRS hyperscanning studies, *eneuro*, 9(2), 2022.

[3] You, J., Jung, M., Shin, Y., & Kim, K., Teacher-Student Inter-Brain and Behavioral Synchronies in Remote Education, *IEEE Access*, 2025.

[4] You, J., Jung, M., & Kim, K., Can People's Brains Synchronize during Remote AR Collaboration?, *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[5] You, J., Jung, M., & Kim, K. K., Inter Brain Synchrony in Remote AR Education: Can Warming up Activities Positively Impact Educational Quality?, *IEEE International Symposium on Mixed and Augmented Reality*, 584-593, 2024.

[6] Lindell, A. K., In your right mind: Right hemisphere contributions to language processing and production, *Neuropsychology review*, 16(3), 131-148, 2006.

효과적 인간-AI 페인팅 협업을 위한 VLM 에이전트 기반 비평시스템*

류보경[0], 김영준
이화여자대학교 컴퓨터공학과
{rbogyeong, kimy}@ewha.ac.kr

VLM Agent-Based Critique System for Effective Human-AI Painting Collaboration

BoGyeong Ryu [0], Young J. Kim
Dept. of Computer Science and Engineering, Ewha Womans University

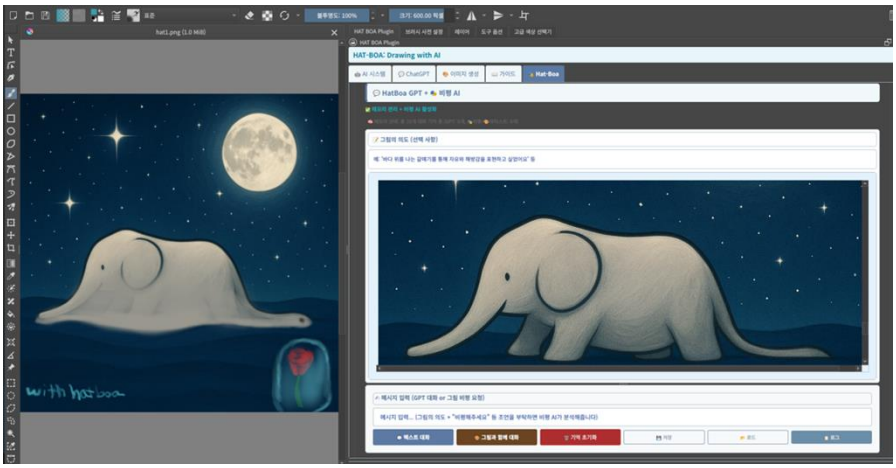


Figure 1: Human-AI collaborative art creation system that analyzes the artwork (left) and provides enhanced interpretations (right) through multi-agents.

Abstract

We propose a multi-agent system for interactive art creation through human-AI collaboration. The system allows users to receive feedback from a Vision-Language Model based on their canvas content and input intentions, as well as to access AI-generated painting suggestions resulting from this feedback. Agents are consecutively executed to manage feedback generation, artistic planning, and prompt formulation for image editing models. The system is implemented as a Krita plugin, allowing users to incorporate AI-generated results as needed into their creative processes. The pipeline can be repeated until the canvas is refined to the user's satisfaction.

1. Introduction

In the novel 'The Little Prince' by Antoine de Saint-Exupéry, the boy drew a boa constrictor, but adults perceived it as a hat, as illustrated in Figure 2. "My drawing was not a picture of a hat. It was a picture of a boa constrictor digesting an elephant. I gave up what might have been a magnificent career as a painter." – This excerpt illustrates how misunderstood artistic intentions can discourage creative expression when proper feedback is absent. Contemporary AI art creation encounters similar communication challenges. Current generative AI systems focus on immediate output generation without artistic critique or progressive feedback mechanisms.

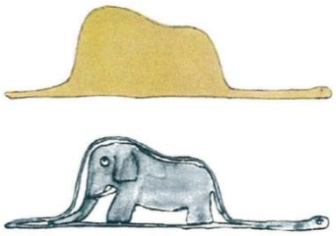
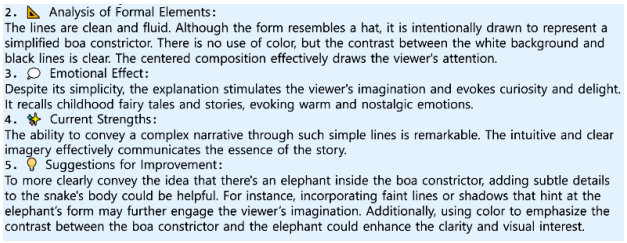


Figure 2: A drawing of a boa swallowing an elephant

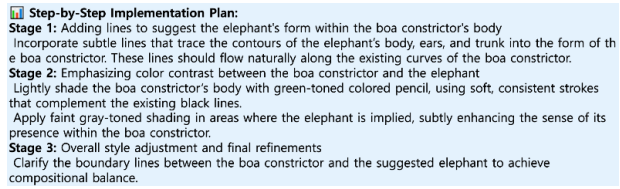
This paper presents a multi-agent system with Critique, Artist, and Prompt agents that provides artistic critique and enhancement while preserving original creative visions.

* 구두 발표 초록 논문 (extended abstract)
* 본 연구는 ITRC/IITP프로그램 IITP-2025-RS-2020-II201460 와 연구재단 2022R1A2B5B03001385의 지원으로 수행되었음

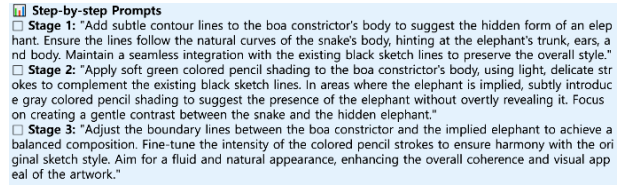
2. Multi-Agent System for Art Collaboration



(a) Critic Agent



(b) Artist Agent



(c) Prompt Agent

Figure 3: Multi-Agent System Components and Results

2.1. Critique Agent

Users can receive artistic critique from a VLM-based critique agent. The system examines user text input to identify keywords, activating the critique agent. The critique agent analyzes the canvas image and the artistic intent provided by the user. The agent analyzes composition, visuals, and emotion to suggest improvements aligned with the user's intent.

2.2. Critique-Based Planning and Image Editing

The artist agent incorporates the user's canvas, creative intent, and critique analysis through sequential reasoning to formulate a three-stage improvement plan. The prompt agent converts this plan into prompts for image editing models such as InstructPix2Pix [1] and GPT-Image-1 [2].

2.3. Painting Interface

Exploiting the Krita AI Diffusion plugin [3], we implement our system as a Krita plugin that enables multi-agent execution. Users can paint on the canvas and input their creative intent, activating the critique-planning-editing workflow.

Caption Type	Pre-critique Canvas-Caption Alignment	Post-critique Canvas-Caption Alignment	Δ Score	Improved Cases
First Caption	0.349	0.476	+0.127	10/10
Second Caption	0.194	0.492	+0.298	10/10

Table 1: VQA score [4] comparison before and after critique-based revision (mean of ten trials)

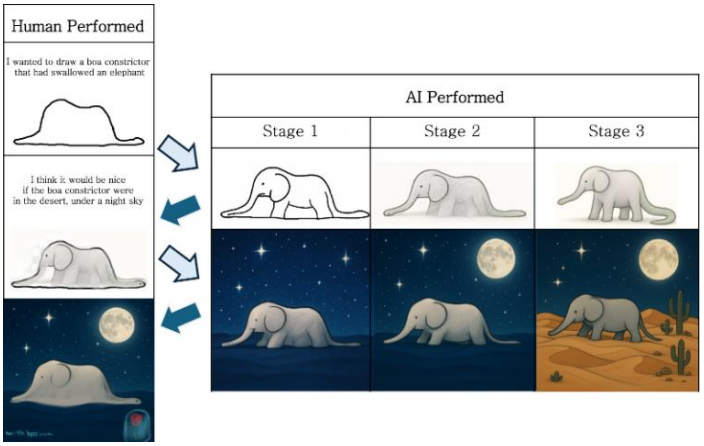


Figure 4: Human-AI Collaborative Creation Results. In the left column, the first prompt drew an elephant from an outline to a shaded image; the second one depicted a desert night with moon and cacti; finally, the user finished painting with a rose and a signature.

3. Experimental Results

We demonstrate our system through boa constrictor drawing experiments (Figures 3, 4). Across multiple iterations, the Critique, Artist, and Prompt agents collaborated with the user to improve the canvas while preserving their original intent. VQA evaluation [4] (Table 1) shows that post-critique canvases consistently align better with critique-driven captions than pre-critique versions, validating the system's feedback effectiveness.

4. Conclusion

We presented a multi-agent system that leverages VLM for art critique and editing, demonstrating a novel human-AI collaborative creative interface. The system maintains artistic intent while enabling progressive visual improvement through human-AI interaction. Future work includes adding voice input for real-time critique, reference search, and idea generation, enabling more fluid human-AI collaboration.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros, InstructPix2Pix: Learning to Follow Image Editing Instructions, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.
[2] OpenAI, Introducing our latest image generation model in the API, https://openai.com/index/image-generation-api/, 2025.
[3] Acly, Krita-AI-Diffusion, https://github.com/Acly/krita-ai-diffusion, 2023.
[4] Zhiqiu Lin, Deepak Pathak, Baiqi Li, et al, Evaluating Text-to-Visual Generation with Image-to-Text Generation, *2024 European Conference on Computer Vision (ECCV)*. Springer, 2024.

논문 발표

애니메이션

사전 렌더링 캐릭터 애니메이션을 위한 경로 기반 캐릭터 모션 조작 시스템

이지원 [0], 이윤상
한양대학교 컴퓨터소프트웨어학과
{babap8514, yoonsanglee}@hanyang.ac.kr

Neural Motion Path: Motion Path-based Manipulation System for Pre-rendered Character Animation Authoring

Jiwon Yi[0], Yoonsang Lee
Department of Computer Science, Hanyang University

요약

본 연구에서는 캐릭터의 전신 모션 경로를 조작하여 캐릭터 애니메이션을 합성하는 시스템을 제안한다. 학습된 모델을 통해 사용자는 경로의 세부 사항을 일일이 조정하지 않고도 상황에 적합한 사실적인 모션을 생성할 수 있다.

1. 서론

고품질 캐릭터 애니메이션 제작은 영화, TV 시리즈와 같은 사전 렌더링 형식의 콘텐츠 제작에서 필수적인 작업이다. 이러한 작업은 반복적으로 이뤄지며, 잦은 수정이 필요하고 즉각적인 시각적 피드백을 요구한다.

우리는 이러한 저작 과정을 지원하기 위한 딥러닝 기반 저작 시스템 Neural Motion Path (NMP)를 제안한다. NMP 는 관절 단위의 모션 경로를 조작함으로써 전신 모션을 편집할 수 있게 한다. NMP 는 모션 생성기와 제약 조정기를 결합한 구조로 되어 있으며, 각 모듈을 통해 사용자의 세부 조정 없이 입력 경로에 부합하는 모션을 자동으로 생성하고 간단한 경로 편집을 통해 모션을 정밀하게 조정할 수 있도록 한다. 결과적으로 초보자도 쉽게 익힐 수 있으면서 정교한 애니메이션 저작도 가능하게 한다.

우리의 연구에서는 간단하고 직관적인 모션 경로 편집을 위한 조작 기능들을 구현하여 사용자가 보다 간단하고 쉽게 경로를 조작할 수 있도록 했다. 또한 널리 사용되고 있는 애니메이션 저작 도구인 Blender 의 애드온으로 구현하여 효율적이고 사용자 친화적인 환경을 제공하는 동시에 전문적인 작업에도 무리 없이 통합될 수 있음을 보인다.

* 구두(포스터) 발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회 (논문지)에 제출 중임.
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2023-00222776) 및 문화체육관광부 및 한국콘텐츠진흥원의 2024 년도 문화체육관광 연구개발사업 지원을 (RS-2024-00399136) 받아 수행되었음

2. Neural Motion Path 개요

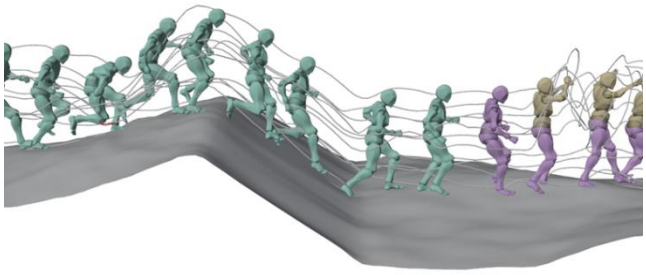


그림 1: 모션 원본/편집/합성 예시

사용자가 모션 클립을 불러오면 시스템은 전신의 모션 경로를 추출하여 화면에 표시한다. 이후 사용자는 다양한 조작 기능을 사용해 전신 모션 경로를 생성한다. NMP 시스템은 만들어진 전신 모션 경로를 입력으로 받아 그에 대응하는 모션을 생성한다.

3. 모션 생성기

우리의 모션 생성기는 [1]에서 영감을 받아 설계되었으며, 다음과 같은 구조를 갖는다.

$$MG(M_i^{ctx}, R_i^{ctx}, P_i) \rightarrow (P_{i+1}, \Delta T_i)$$

전신 모션 경로로부터 i 번째 프레임을 기준으로 전후 1 초간 샘플링한 문맥 모션 경로 M_i^{ctx} 와 현재 포즈 P_i를 받아, 다음 프레임의 포즈 P_{i+1}과 캐릭터 좌표계 갱신값 ΔT_i를 출력한다.

일반적인 경우와 달리 인간의 포즈를 완전하게 표현하려면 회전 정보가 필수지만 이를 모션 경로에 포함하면 조작이 복잡해진다. 따라서 M_i^{ctx}는 회전 정보 없이 위치 정보만 포함한다. 그러나 모션 생성기 내의 회전 네트워크가 위치 및 그에 따른 속도 정보로 이루어진 M_i^{ctx}로부터 회전 값 R_i^{ctx}를 예측하며, 포즈 네트워크는 M_i^{ctx}, R_i^{ctx}, P_i를 이용해 다음 프레임의 포즈를 생성한다. 이러한 위치 기반 설계를 통해 직관적인 위치 기반 편집 인터페이스를 유지하면서도 모션의 표현력을 강화할 수 있다.

회전 네트워크는 [2]와 같이 파라미터 기반의 Mixture of Experts (MoE) 구조를 사용한다. 게이팅 네트워크는

M_i^{ctx}에서 추출한 관절 속도 궤적을 입력으로 받아 혼합 가중치를 출력한다. 포즈 네트워크 또한 회전 네트워크와 동일한 MoE 구조를 갖는다.

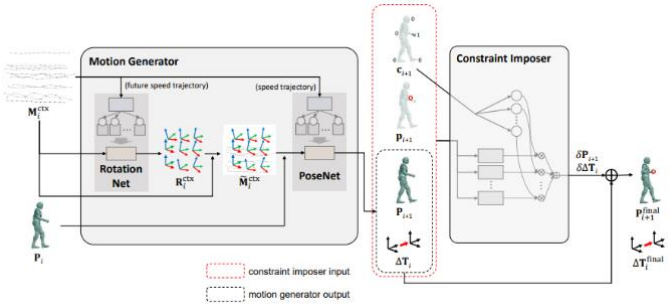


그림 2: 네트워크 Overview

4. 제약 조정기

모션 생성기는 모션 경로를 세밀하게 조정하지 않아도 상황에 적절한 동작을 생성한다. 이로 인해 개별 관절이 사용자가 지정한 경로를 정확히 따르지 않을 수도 있다. 하지만 애니메이션 저작에서는 정밀 제어가 자주 요구되기 때문에 우리는 제약 조정기를 도입하여 출력 모션이 사용자가 지정한 말단 관절의 경로상의 제약을 정확히 따르도록 유도한다.

$$CI(c_{i+1}, p_{i+1}, P_{i+1}, \Delta T_i) \rightarrow (\delta P_{i+1}, \delta \Delta T_i)$$

이 모듈은 기존 자세와 각 말단 관절에 대한 제약 강도 c_{i+1} 및 목표 위치 p_{i+1}를 입력으로 받아 조정된 자세를 출력한다. c_{i+1}는 0~1 의 값으로 정의되며, 부드러운 전환이 가능해 제약 구간과 비-제약 구간 사이의 자연스러운 연결을 지원한다. 출력 값의 경우 기존 포즈와 좌표계 갱신 값에 더해질 조정 값의 형식으로 출력된다.

제약 조정기는 우리가 제안하는 Explicit-Weight Sparse Expert Model (EW-SEM)로 이루어져 있다. 이 구조는 말단 관절과 동일한 개수의 전문가 네트워크를 사용한다. 학습 시 각 전문가의 출력은 대응하는 말단 관절의 제약 강도에 따라 가중치가 부여되며, 이를 통해 각 전문가가 해당 목표 위치를 만족시키는 조정 자세를 학습하도록 유도된다.

우리는 [3]의 접근 방식을 따라 기존 자세를 보존하는 학습 방식을 채택하였다. 기존 자세를 입력으로 제공하게 되면 정보량이 충분해져 더 높은 예측 가능성을 갖게 되고 프레임 단위로 생성된 포즈 사이의 일관성을 가지게 된다. 손실 함수의 경우 말단 관절은 제약 목표 위치에 가깝도록, 그 외의 관절은 기존 자세를 보존하도록 구성되어 있다.

5. 조작 기능

NMP 는 사용자가 임의로 관절과 시간 구간을 선택해 모션 경로를 편집 및 조작할 수 있도록 하는 기본적인 **편집** 기능을 제공한다. 추가로 전신 모션 경로를 **분할**하고 **연결**하여 새로운 모션 시퀀스를 생성하거나, 원본 모션 경로의 일부 관절 모션 경로를 다른 모션 경로에 덮어씌우는 식으로 두 전신 모션 경로를 **혼합**하거나, 모션 경로를 다른 지형 위로 **적응**시키거나, 선택된 주기 동작을 **복제**하여 사용자가 그린 경로를 따라 움직이는 모션을 생성하거나, 선택한 말단 관절 임의의 구간의 **제약**을 설정 및 해제하는 기능을 제공한다.

또한 제약을 이용해 자동 접촉 보존(Automatic Contact Preservation, ACP) 기능을 제공한다. 이는 원본 모션에서 지면에 접촉한 시점의 발에 제약 강도 1 을 부여해 발의 위치가 접촉 지점에 유지되도록 한다. 이는 편집 후 접촉 타이밍의 발 미끄러짐을 완화하는 역할을 한다.

이러한 기능들은 Blender 애드온 형태로 구현되어 프레임 단위의 인터랙티브 모션 생성을 지원한다.

6. 결론

우리는 사전 렌더링 캐릭터 애니메이션 저작을 위한 시스템 NMP 를 제안한다. 이 시스템은 사용자가 경로를 편집하고 제약을 추가함으로써 전신 모션을 생성할 수 있도록 한다. NMP 는 문맥을 고려한 모션 생성을 위한 모션 생성기와, 정확한 말단 관절 정렬 및 발 미끄러짐 방지를 위한 EW-SEM 을 사용하는 제약 부여 모듈을 결합하여 구현되었다. 제약 부여 모듈 구조 자체는 프레임 단위로 동작하므로 시간적 일관성을 보장하지 않지만, 포즈 보존을 통해 연속적인 편집 과정에서도 부드러운 조정이 가능하게 한다. 경로를 이루는 점의 위치를 조작하는 지금의 편집 방법의 특성 상 정밀하거나, 부드러운 편집을 하는 것은 초보 사용자에게는 조금 어려울 수 있다. 따라서 편집의 사용성을 높이기 위해 베지어 기반 경로 보간 방식[4]의 도입해보는 것도 좋은 방법이다. 또한 과도한 제약은 준비되지 않은 갑작스러운 점프나 몸이 비현실적으로 기우는 등 부자연스러운 결과를 유발할 수 있는데, 제약 부여 모듈을 물리 기반 기법과 결합하는 것이 이러한 문제를 완화하는 데 도움이 될 수 있으며, 매우 유망한 발전 방향이다.

참고문헌

[1] Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura, Neural Animation Layering for Synthesizing Martial Arts Movements, *ACM Trans. Graph., Vol. 39, No. 4*, 2021

[2] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito, Mode-adaptive neural networks for quadruped motion control, *ACM Trans. Graph., Vol. 37, No. 4*, 2018

[3] Agrawal, Dhruv, Guay Martin, Buhmann Jakob, Borer Dominik, and W. Sumner Robert, Pose and Skeleton-aware Neural IK for Pose and Motion Editing, *SIGGRAPH Asia 2023 Conference Papers, Article No. 4, pp. 1-10,* 2023.

[4] Studer Justin, Agrawal Dhruv, Borer Dominik, Sadat Seyedmorteza, W. Sumner Robert, Guay Martin, and Buhmann Jakob, Factorized Motion Diffusion for Precise and Character-Agnostic Motion Inbetweening, ACM SIGGRAPH Conf. Motion, Interaction, and Games (MIG '24), Article No.11, pp. 1-10, 2024.

PhysicsFC: 물리 기반 축구 선수 컨트롤러를 위한 사용자 제어 스킬 학습

김민수 [0], 정은호, 이윤상
한양대학교 인공지능학과, 한양대학교 컴퓨터 소프트웨어학과, 한양대학교 컴퓨터 소프트웨어학과
igotaspot426@gmail.com, jho6394 @hanyang.ac.kr, yoonsanglee@hanyang.ac.kr

PhysicsFC: Learning User-Controlled Skills for a Physics-Based Football Player Controller

Minsu Kim[0], Eunho Jung, Yoonsang Lee
Dept. of Artificial Intelligence, Hanyang University, Dept. of Computer Science, Hanyang University

요약
PhysicsFC는 사용자 입력에 따라 물리 기반 축구 캐릭터가 드리블, 트래핑, 이동, 킥 등의 기술을 자연스럽게 수행하고 전환할 수 있도록 학습된 제어 시스템이다. 각 기술은 별도의 정책으로 구성되어 있으며, 물리 기반 모션 임베딩 모델을 활용해 모션 재현이 가능하다. 기술 간 전환을 부드럽게 하기 위해 FSM 기반 제어기와 전이 초기화 기법(STI)을 적용하였다. 본 시스템은 다양한 시나리오에서 사용자 조작에 따라 현실감 있는 축구 플레이를 구현할 수 있음을 보였다.

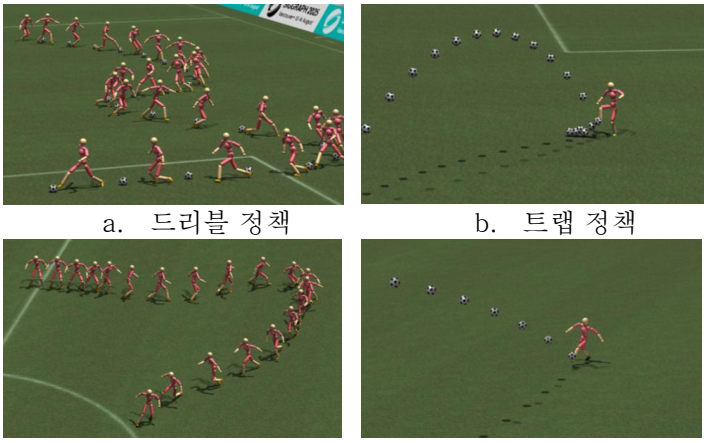
1. 서론
물리 시뮬레이션 기반의 축구 캐릭터 제어는 사실적인 공과의 상호작용을 구현하는 데 어려움이 있다. 기존 게임은 캐릭터는 키프레임 애니메이션으로, 공만 물리로 처리하여 비현실적인 동작이 발생한다. 이전 연구들에서 물리환경에서 축구 시뮬레이션 하는 사례들이 있었지만 유저 입력을 처리하기 어려웠거나 [1], 부자연스러운 모션 [2], [3] 을 보여주었다. PhysicsFC는 사용자 입력에 따라 드리블, 트래핑, 킥 등 다양한 기술을 물리 기반으로 자연스럽게 수행할 수 있게 한다. 기술별 정책은 모션 임베딩 모델을 활용해 학습되며, FSM 구조와 특수한 초기화를 통해 기술 간 부드러운 전환을 지원한다. 이로써 실제 축구에 가까운 물리적 상호작용과 조작 가능한 플레이를 구현한다.

* 구두 발표논문, 요약논문 (Extended Abstract)
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 SIGGRAPH 2025 에서 발표 및 ACM Transactions on Graphics에 게재 예정임.
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2023-00222776) 과 정보통신기획평가원의 지원(No.RS-2020-II201373,인공지능대학원지원(한양대학교)) 과 문화체육관광부 및 한국콘텐츠진흥원의 2025 년도 문화체육관광 연구개발사업 지원을 (RS-2024-00399136) 받아 수행되었음.

2. 개요
우리의 방법은 물리 기반 축구 캐릭터가 다양한 축구 스킬을 사용자 입력에 따라 수행할 수 있도록 설계되었다. 저수준 정책은 CALM[4] 모션 임베딩 모델을 활용해 실제 축구 모션 데이터를 물리 시뮬레이션 환경에서 재현 가능하도록 사전 학습하였다. 저수준 정책은 잠재변수 Z 를 입력받아 그에 해당하는 동작을 출력한다. 각각의 축구 스킬(드리블, 트래핑, 이동, 킥)은 앞서 사전학습한 CALM[4] 기반의 저수준 정책 위에 개별적인 고수준 스킬 정책을 계층적으로 구성하여 학습되었다. 각 고수준 정책은 각 스킬의 목표를 입력받아 해당 목표를 달성할 수 있는 잠재변수 Z 를 출력한다. 이러한 계층적 구조는 사용자 의도에 따라 고수준 정책이 목표를 설정하면, CALM[4]이 이에 대응하는 자연스러운 모션을 생성하는 방식이다. 기술 간 전환은 FSM(유한상태기계) 구조를 통해 정의되며, 상황 및 입력에 따라 적절한 기술 정책으로 전환된다. 또한, 각 스킬 정책은 FSM 내 선행 기술로부터 초기 상태를 샘플링하는 Skill Transition-Based Initialization (STI) 방식으로 학습되어 전환의 부드러움을 확보하였다. 이러한 전체 구조는 자연스럽고 물리 기반의 축구 기술을 조합하여 사용자 상호작용 중심의 플레이가 가능하도록 한다.

3. 축구 정책 학습
우리 시스템은 드리블, 트래핑, 이동, 킥의 4 가지 축구 스킬에 대해 각각 고유한 정책을 학습하였다. 드리블은 공을 목표 속도 방향으로 유지하며 전진하는 것이 목표이며, 공의 속도, 위치, 캐릭터의 진행 방향을 기준으로 보상을 부여한다. 초기화는 트래핑 또는 이동 상태에서 전이된 상태를 사용한다. 트래핑은 공을 지정된 신체 부위로 안정적으로 제어하는 것이 목표이며, 충돌 전에는 공과 신체의 거리, 충돌 후에는 공과 몸의 상대 속도를 기준으로 보상이 주어진다. 초기화는 이동 상태에서 공이 날아오는 상황을 시뮬레이션해 구성한다.

이동은 사용자가 지정한 속도와 방향으로 움직이며, 다양한 움직임 스타일(전진, 후진, 측면 등)을 자연스럽게 재현하는 것이 목표이다. 다양한 보행 모션을 효과적으로 활용하도록 학습하기 위해, 모션 데이터에 내재된 목표잠재 표현 쌍을 활용하는 Data-Embedded Goal-Conditioned Latent Guidance (DEGCL) 기법을 통해모션 다양성과 목표 정합성을 동시에 확보하며, 초기화는 항상 정지 상태에서 시작된다. 킥은 공을 지정된 방향과 속도로 차는 것이 목표이며, 공의 초기 속도와 목표 속도의 차이에 기반한 보상을 받는다. 초기화는 드리블 중 상태 또는 정지 상태에서 시작된다. 모든 스킬은 CALM[4] 기반 저수준 제어기를 latent vector로 유도하는 방식으로 구현되며, 전이된 초기 상태에서 빠르게 안정된 동작을 수행할 수 있도록 설계되었다.



a. 드리블 정책 b. 트랩 정책

c. 이동 정책 d. 킥 정책

그림 1: 축구 스킬 정책 예시

4. 실험 및 데모

실험에서는 제안한 STI의 효과를 검증하였다. STI를 적용한 경우, 스킬 전환 직후 목표 달성까지 걸리는 시간이 크게 단축되었고, 성공률 또한 높게 나타났다. 예를 들어 Trap→Dribble 전이에서는 STI 적용 시 드리블 목표 도달 시간이 약 77% 빨라졌으며, Kick 전이에서는 비적용 모델이 거의 킥에 실패했다. 이는 각 스킬이 전이 직후의 상황에 빠르게 적응하도록 학습되었음을 의미한다.

또한 다양한 상호작용 데모를 통해 시스템의 실제 활용 가능성을 확인하였다. 사용자가 한 명의 캐릭터를 조작하여 드리블, 패스, 슛을 수행하거나, 팀 플레이를 통해 '원투패스(give-and-go)'를 구현할 수 있다. 상대 캐릭터와의 경쟁 상황에서 트래핑과 드리블을 반복하거나, 11 vs 11 의 축구 경기를 시뮬레이션하는 것도 가능하다. 모든 상황에서 제안된 FSM 구조와 스킬 정책이 실시간으로 연동되어, 자연스러운 물리 기반 축구 플레이를 제공한다.

그림 1: 유저 컨트롤 되는 11-11 축구경기 데모

5. 결론

본 연구에서는 CALM[4] 기반 모션 임베딩과 계층적 스킬 정책, FSM 구조, 그리고 STI 기법을 결합하여 물리 시뮬레이션 환경에서 사용자 조작 가능한 축구 캐릭터 제어 시스템인 PhysicsFC를 제안하였다. 이를 통해 자연스럽고 민첩한 축구 기술 수행과 기술 간 전환이 가능함을 정량적 실험과 다양한 데모를 통해 확인하였다. 다만, 현재 드리블과 킥이 특정 발에 편중되거나, 마그누스 효과 등 실제 공역학을 반영하지 못한 한계가 존재한다. 추후 연구에서는 양발 사용 학습, 고급 스킬 다양화, 수비 및 충돌 상호작용 확장, 더 정교한 공 물리 모델 적용 등을 통해 보다 사실적인 축구 플레이를 구현할 수 있을 것으로 기대된다.

참고문헌

[1] Seokpyo Hong, Daseong Han, Kyungmin Cho, Joseph S. Shin, and Junyong Noh, Physics-based full-body soccer motion control for dribbling and shooting. ACM Trans. Graph. 38, 4, Article 74 (August 2019), 12 pages, 2019

[2] Xue Bin Peng, Glen Berseth, Kangkang Yin, and Michiel Van De Panne, DeepLoco: dynamic locomotion skills using hierarchical deep reinforcement learning. ACM Trans. Graph. 36, 4, Article 41 (August 2017), 13 pages, 2017

[3] S. Liu, G. Lever, Z. Wang, J. Merel, S. M. A. Eslami, D. Hennes, W. M. Czarnecki, Y. Tassa, S. Omidshafiei, A. Abdolmaleki, N. Y. Siegel, L. Hasenclever, L. Marris, S. Tunyasuvunakool, H. F. Song, M. Wulfmeier, P. Muller, T. Haarnoja, B. D. Tracey, K. Tuyls, T. Graepel, and N. Heess, From motor control to team play in simulated humanoid football, Science Robotics, vol. 7, no. 69, 2022

[4] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Pen, CALM: Conditional Adversarial Latent Models for Directable Virtual Characters. In ACM SIGGRAPH 2023 Conference Proceedings (SIGGRAPH '23). Association for Computing Machinery, Article 37, 1–9, 2023

물리 시뮬레이션 캐릭터의 모션 학습을 위한 적응형 부위별 잠재 토큰*

배진석, 이영환, 임동근, 김영민
서울대학교 전기·정보공학부
{capoo95, frredy99, rms2836, youngmin.kim}@snu.ac.kr

PLT: Part-Wise Latent Tokens as Adaptable Motion Prior for Physically Simulated Characters

Jinseok Bae, Younghwan Lee, Donggeun Lim, Young Min Kim
Dept. of Electrical and Computer Engineering, Seoul National University

요약

물리 시뮬레이션과 데이터 기반 애니메이션 연구에서는 캐릭터가 모션 캡처 데이터를 모방함으로써 자연스러운 전신 동작을 생성할 수 있지만, 기존 방식은 데이터셋 범위 밖의 다양한 상황에서 효율적인 적응이 어렵다는 한계가 있다. 본 논문에서는 이러한 한계를 극복하기 위해 신체 부위별 움직임을 독립적으로 표현하고 조합할 수 있는 잠재 토큰 기반 정책 구조를 제안하며, 부위별로 생성된 움직임을 안정화하기 위한 추가 신경망 구조를 함께 도입한다. 다양한 환경에서의 실험을 통해 제안한 방법론이 안정적이고 우수한 작업 성능을 나타냄을 확인하였으며, 특정 신체 부위의 움직임만을 선택적으로 조정하여 환경 변화에 더욱 효과적으로 대응할 수 있는 부위별 적응 방식을 제시한다.

1. 서론

물리 기반 애니메이션은 모션 캡처 데이터를 활용하여 사실적이고 자연스러운 캐릭터 모션을 생성하는 데 크게 기여해 왔다. 특히 최근 연구에서 도입된 계층적 제어 구조[1,2]는 방대한 모션 데이터셋에서 모션에 대한 잠재 공간에서의 사전지식(latent motion prior)을 학습하고 이를 통해 다양한 작업에 적합한 제어 정책을 효율적으로 도출하는 데 성공하였다. 하지만 이러한 접근 방식은 새로운 모션 스킬의 발굴과 데이터셋 밖의 상황에 대한 작업 적응성에서 여전히 한계를 보이고 있다. 이는 기존의 모션 데이터셋이 일상적인 인간 동작의 모든 범위를 포괄하기 어려워 발생하는 문제이다.

기존 연구들 중 일부는 서로 다른 데이터 소스로부터 신체 부위별 움직임을 조합하여 데이터 효율성과 모션 다양성을 향상시키는 방식을 제안하였다[3]. 그러나 이 접근 방식들은 부위 간 움직임의 호환성과 상호작용을 충분히 고려하지 않았고, 특정한 시나리오 외에는 범용적으로 재사용하기 어려웠다는 한계를 가진다. 따라서 본 논문에서는 부위별 움직임의 호환성을 높이고 다양한 환경에 효율적으로 적응 가능한 정책을 위해 이산 잠재 토큰 (latent token) 기반의 새로운 정책 아키텍처를 제안한다. 구체적으로, 신체 부위별로 독립적인 코드북(codebook)을 구축하고, 이를 통해 각 부위의 전문화된 모션 스킬을 이산 잠재 토큰 형태로 표현하여 구조적으로 모션을 분해하고 조합하는 방식을 제시한다.

2. 이산적 잠재 모델 (Discrete Latent Model)

본 문단에서는 본 연구의 기반이 되는 이산 잠재 모델(discrete latent model)[2]에 대해 설명한다. 계층적 제어 파이프라인은 대규모 모션 데이터셋에서 학습된 모션 스킬을 효율적으로 재사용할 수 있는 실용적인 프레임워크를 제공한다. 파이프라인은 두 단계로 구성된다. 첫째, 모방 학습 단계에서는 방대한 모션 데이터 컬렉션으로부터 모션 사전지식의 이산 잠재 공간을 구축한다. 둘째, 작업 학습 단계에서는 획득된 잠재 표현을 기반으로 계층적 구조를 활용하여 작업 특화된 강화학습 정책을 훈련한다.

계층적 제어에서 이산적인 액션 공간의 활용은 강화학습 과정에서 긍정적인 효과를 제공할 수 있다. 기존의 연속 벡터 기반 잠재 모션 임베딩은 사후 붕괴(posterior collapse) 문제로 인해 하위 작업들의 효율성을 저하시킬 가능성이 있다. 반면, 이산 잠재 벡터는 이러한 문제를 자연스럽게 회피하며 표현의 품질을 유지할 수 있다. 이는 액션 선택의 복잡성을 줄이고 탐색 및 최적화를 간소화하여 보다 효과적인 학습을 가능하게 한다.

3. 부위별 잠재 토큰 (Part-Wise Latent Tokens)

그림 1에서와 같이, 본 연구의 정책 함수 아키텍처는 잠재 모션 사전지식을 부위별 구성요소로 구조적으로 분해한다. 이러한 설계는 에이전트가 개별 신체 부위의 움직임 특성에 집중하여 효과적인 잠재 토큰을 학습하도록 유도한다. 우선, 에이전트의 신체를 K개의 그룹으로 나누고(예를 들어, K=2의 경우 상체와 하체로 구분), 잠재 벡터를 각 그룹에 따라 K개의 공간적 세그먼트로 나눈다.

이산 잠재 모델을 기반으로 확장된 본 정책 아키텍처는

---

* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 SIGGRAPH 2025에 발표될 예정임.

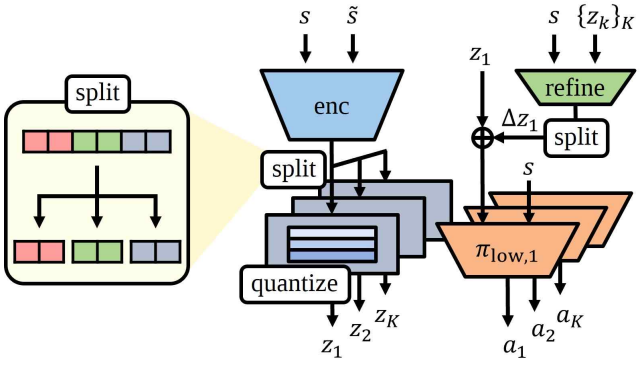


그림 1: 부위별 잠재 토큰 및 정제 네트워크

K개의 코드북과 각 코드북에 대응하는 저수준 정책(low-level policies)을 사용한다. 모방 학습 단계에서 인코더는 출력되는 잠재 벡터를 K개의 세그먼트로 나누고, 각 세그먼트는 해당 코드북 내에서 가장 유사한 잠재 토큰으로 매핑된다. 양자화된 이후, 저수준 정책은 선택된 잠재 토큰과 전신 상태를 바탕으로 개별 신체 부위의 액션을 결정한다.

이산 잠재 모델을 확장하여, 우리의 정책 아키텍처는 K개의 코드북과 해당하는 저수준 정책들을 사용한다. 모방 학습 단계에서 인코더는 출력 y를 K개 세그먼트로 분할한다. 각 연속 잠재 벡터는 해당 코드북에서 가장 가까운 잠재 토큰에 매핑된다. 양자화 후, 저수준 정책은 잠재 토큰과 전신 상태를 기반으로 신체별 액션을 예측한다.

부위별 잠재 토큰(PLT)에 더하여, 본 논문에서는 이산 잠재 벡터를 연속적인 정제 벡터로 보완하여 전신 움직임의 일관성을 유지하는 정제 네트워크(Refinement Network)를 제안한다. PLT 방식은 모션의 다양성과 작업 성능을 향상시키지만, 개별 부위의 움직임을 단순히 조합하는 것만으로는 전신의 전역적 조정이 부족할 수 있다. 이를 보완하기 위해 정제 네트워크는 연결된 잠재 토큰들과 현재의 신체 상태를 입력으로 받아 연속 벡터를 출력하며, 이 벡터는 각 부위 잠재 토큰에 대한 세부적 조정을 위한 오프셋(offset) 형태로 세그먼트화되어 적용된다.

4. 실험 결과

4.1. 모방 성능

본 연구에서 제안된 방법론은 LaFAN1[4], AMASS[5]와 같은 대규모 모션 캡처 데이터셋에서 향상된 모방 제어 (imitation control) 성능을 보인다. 구체적으로, 기존의 계층적 구조 모델들에 비해 정성적, 정량적 평가에서 모션 추적 (tracking) 성능이 향상되었으며, 특히 학습 과정 중 한 번도 보지 않은 새로운 분포의 모션들에 대해 안정적인 모방 능력을 보였다. 특히, 제안된 PLT 모델은 밸런스 유지가 어려운 다이나믹한 모션들에 대해 안정적으로 모방을 수행할 수 있었다.

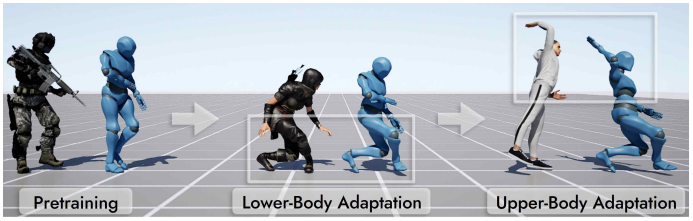


그림 2: 단계별로 각 부위별 움직임을 적응시킨 결과

4.2. 작업 성능

제안된 PLT 구조는 고수준 정책 (high-level policy)의 학습에 높은 효율성을 보인다. 본 연구에서는 제한된 개수의 트래커를 추적하는 N-body Tracking을 비롯해, 에이전트의 Actuator 일부가 데미지가 입혀진 상황에서의 네비게이션 작업 등 도전적인 8개의 상황에서의 작업 수행 능력을 측정하였다. 그 결과, 모든 시나리오에서 기존 연구들에 비해 평균적으로 높은 보상 (reward)을 받는 것을 확인할 수 있었다. 특히 밸런스 유지가 핵심적인 상황에서는 정제 네트워크의 존재가 PLT 에이전트의 안정성에 큰 효과를 보이는 것을 실험적으로 증명하였다. 또한 그림 2에서 나타난 것과 같이, 기존의 연구들과는 다르게 특정한 부위의 모션 사전지식을 업데이트하는 부위별 적응 실험을 통해 여러 모션 데이터셋을 동적으로 조합하는 것이 가능함을 보였다.

5. 결론

부위별 잠재 토큰에 더하여, 본 논문에서는 이산 잠재 벡터를 연속적인 정제 벡터로 보완하여 전신 움직임의 일관성을 유지하는 정제 네트워크를 제안한다. 이 방식은 개별 부위의 움직임을 유기적으로 조합하여 모션의 다양성과 작업 성능을 강건하게 향상시킨다. 작업에 따라 최적의 분할 전략을 자동으로 학습하는 방식은 향후 주요 연구 방향성이 될 수 있다.

참고문헌

[1] J.Won, D.Gopinath and J.Hodgins, Physics-based character controllers using conditional vaes, *ACM Transactions on Graphics (TOG)*, 41(4):1-12, 2022.

[2] Q.Zhu, H.Zhang, M.Lan, and L.Han, Neural categorical priors for physics-based character control, *ACM Transactions on Graphics (TOG)*, 42(6):1-16, 2023.

[3] J.Bae, J.Won, D.Lim, C.Min and Y.Kim, PMP: Learning to physically interact with environments using part-wise motion priors, *ACM SIGGRAPH Conference Proceedings*, 2023.

[4] FG.Harvey, M.Yurick, D.Nowrouzezahrai and C.Pal, Robust motion in-betweening., *ACM Transactions on Graphics (TOG)*, 39(4):60-1, 2020.

[5] N.Mahmood, N.Ghorbani, NF.Troje, G.Pons-Moll and MJ.Black, AMASS: Archive of motion capture as surface shapes. *InProceedings of the IEEE/CVF international conference on computer vision*, 2019.

단일 2D RGB 영상을 이용한 보행주기 분석 프레임워크 *

김대용[0,1], 신정환[3], 유리[1,2,*]
아주대학교 인공지능학과[1], 아주대학교 소프트웨어학과[2], 서울대학교병원 신경과[3]
so155s@ajou.ac.kr, neo2003@snu.ac.kr, riyu@ajou.ac.kr

GCDN: Detecting Gait Cycle From Monocular Video

Dae-Yong Kim[0,1], Jung-Hwan Shin[3], Ri Yu[1,2,*]
Dept. of Artificial Intelligence, Ajou University[1],
Dept. of Software and Computer Engineering[2],
Dept. of Neurology, Seoul National University College of Medicine[3]

요약

단안 비디오 기반으로 모션 정보를 추출하는 연구[1][2]가 활발히 진행되고 있다. 모션 캡쳐 시스템은 고품질의 모션 데이터를 얻어낼 수 있지만, 접근성이 떨어진다. 반면에, 단안 비디오를 이용하게 되면 별도의 전문적인 장비 없이 스마트폰과 같은 촬영 장비만 있으면 누구나 손쉽게 모션 데이터의 확보가 가능하다. 본 연구는 사람의 다양한 모션 중 보행 분석에 특화된 딥러닝 기반의 프레임워크를 제안한다.

다양한 각도에서 촬영한 2D RGB 영상과 그에 해당하는 3차원 모션 캡쳐 데이터가 포함된 데이터셋[3]을 사용한다. 모션 캡쳐 데이터로부터 발과 지면의 접촉 레이블[4]을 생성하고, Pose Estimator[5]을 통해 영상으로부터 얻은 하체 관절 시퀀스(x, y, confidence)를 좌표계 변환 및 패딩을 거쳐 모델의 입력으로 사용한다. 시퀀스 데이터는 FC layer로 임베딩 후, 트랜스포머의 인코더를 거쳐서 인코더의 출력에 다시 FC layer를 연결해 프레임 단위로 발과 지면 사이의 접촉 정보를 추정을 하는데 이용된다. 위와 같이 학습된 모델을 이용해서 영상의 프레임별 발과 지면 사이의 접촉 정보를 추정하고, 추정된 값을 통해 주요 보행 주기 파라미터를 분석한다. 제안하는 프레임워크는 기존의 모델[6]과 비교했을 때, 보행 지표 분석에서 통계적으로 유의한 성능을 보였다.

1. 서론

본 연구는 사람의 다양한 모션 중에서 특히 보행 분석에 특화된 딥러닝 기반 프레임워크를 제안한다. 기존 보행 분석에 사용되는 장비는 고품질의 데이터를 제공하지만 높은 비용과 전문적인 환경을 요구하기 때문에 접근성이 낮다는 단점이 있다. 이를 극복하기 위해서 단안 비디오만으로 여러가지 보행 파라미터를 추출하는 것이 본 연구의 목표이다.

* 구두발표논문
* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구결과로 수행되었음(IITP-2025-RS-2023-00255968)

보행주기 분석은 세 단계로 구성된다. 첫 번째 단계에서는 모션 캡쳐 데이터와 영상 데이터가 함께 제공되는 데이터셋[3]을 활용하여 네트워크 학습에 적합한 형태로 모델의 학습 데이터셋을 구성한다. 3차원 모션 캡쳐 데이터에서 기존 프레임워크[4]를 이용하여 지면과 발의 접촉 정보를 얻고, Pose Estimator[5]를 통해서 영상에서 2차원 관절 좌표를 추출한다 (그림 1 참조). 두 번째 단계에서는 구축된 데이터셋을 통해 모델을 학습한다.

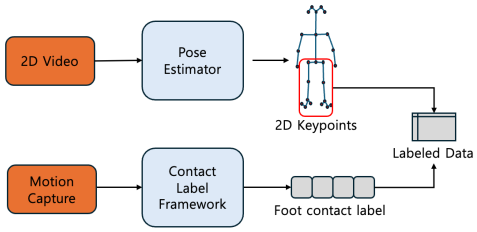


그림 1: 데이터 전처리를 통한 모델 학습 데이터 구성

마지막 단계에서는 학습된 모델로 추정한 발과 지면의 접촉 정보를 이용하여 주요 보행 파라미터(Stance Phase, Swing Phase, Single Support, Double Support)를 분석한다.

해당 프레임워크의 유용성을 효과적으로 평가하기 위해서 학습에 사용되지 않은 실제 의료 현장에서 수집된 데이터를 이용한다. 평가에 사용된 데이터는 서울대학교 병원에서 제공받은 건강인과 파킨슨 환자의 2D RGB 29.97fps 보행 영상 및 압력 센서 매트 데이터를 포함한다. 보행 영상은 동일한 피험자에 대해서 정면과 후면 2가지 각도에서 촬영되었다. 이 데이터를 통해 우리의 프레임워크가 실제 임상 환경에서도 유의미한 보행 분석 결과를 제공하는지 검증한다.

2. 방법

데이터 전처리와 네트워크 학습 그리고 보행지표 분석으로 나누어 기술한다. 데이터 전처리 단계는 네트워크의 입력에 적절한 형태로 데이터를 가공하고 데이터에 대한 레이블을 얻는 것이 목표이다. 네트워크의 학습에서는 데이터 전처리에서 가공된 데이터를 이용하여 트랜스포머 기반 모델의 학습을 진행한다. 모델의 학습 이

후에 학습된 모델로부터 추론한 2D RGB 영상의 각 프레임별 발과 지면 사이의 접촉 정보를 통해서 보행의 주요 파라미터를 추출한다.

2.1. 데이터 전처리

데이터셋[3]에서 보행과 관련된(걷기, 달리기) 동작만을 사용하였다. 모션은 9개의 각도에서 촬영된 영상과 그에 해당하는 모션 캡쳐 데이터가 200fps로 제공된다. 품질 문제로 2번 카메라를 제외한 8개의 카메라 뷰를 사용한다. 영상으로부터 얻은 하체 부분의 2차원 관절 좌표를 골반을 기준으로 상대좌표로 변환한다. 모델의 학습에 필요한 레이블 값을 얻기 위해 기존의 프레임워크[4]를 이용한다. 해당 프레임워크[4]는 모션 캡쳐 데이터를 입력으로 넣으면, 각 프레임별 발과 지면사이의 접촉 정보를 반환한다.

레이블 추정에 사용한 프레임워크[4]의 요구사항에 따라, 짝수 프레임만을 취해 100fps로 다운샘플링 하였다. 총 326, 825 프레임(약 54.5분)을 8:2로 분할하여 훈련에 267,520 프레임(약 44.6분), 검증에 59,309 프레임(약 9.9 분)을 사용한다. 모델 학습 시, 각 영상 시퀀스의 시작점을 랜덤으로 선택하여 랜덤 크롭 방식을 적용한다. 시퀀스 길이가 모델의 입력 길이(임베딩 차원)보다 짧을 경우 제로 패딩을 적용한다.

2.2. 네트워크 학습

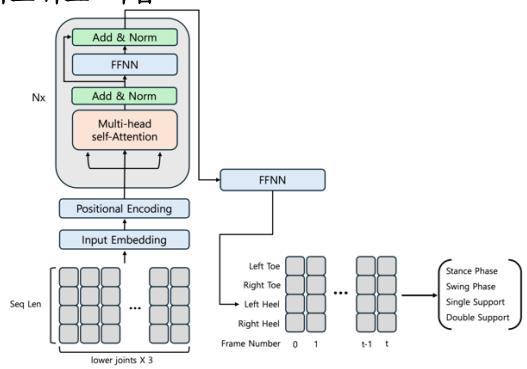


그림 3: 네트워크 구조

지면과 발의 접촉 여부를 판단하는 분류 문제이기 때문에, 트랜스포머 모델에서 인코더의 출력에 완전연결층을 합쳐서 전체적인 모델을 구성한다. 모델의 입력은 골반을 기준으로 변환된 하체 13개 관절의 2차원 좌표와 각 관절의 신뢰도 값을 시퀀스 형태로 구성한 데이터로, 전체 입력 차원은 $\boldsymbol{P} \in \mathbb{R}^{(lower\ joints \times 3) \times seq\_len}$ 이다. 네트워크의 출력은 입력 시퀀스에 대한 각 발의 엄지 발가락과 뒤꿈치의 접촉 여부를 나타내는 이진 값으로, $\hat{Y} \in \mathbb{R}^{4 \times seq\_len}$ 의 형태를 가진다. (그림 3 참조). 모델의 입력 데이터에 시간적 순서 정보를 부여할 수 있게 학습이 가능한 위치 인코딩 방식을 적용한다.

2.3. 보행지표 분석

모델의 유효성을 평가하기 위해서, 오른발 기준으로 보행의 각 단계를 분석 후(그림 4 참조), 해당 정보를 바탕으로 stance phase와 swing phase, single support, double support를 계산하였다. 우리의 모델과 기준으로 설정한 모델[6]의 결과를 정량적으로 비교하기 위해, 서울대학교병원에서 제공받은 18명의 피험자에 대한 압력 센서 매트 결과값의 평균 수치를 비교한다.

3. 결과

표 1: 성능 평가표

Method	Stance % Of Cycle R(Mean)	Swing % Of Cycle R(Mean)	Single Supp % Cycle R(Mean)	Double Supp % Cycle R(Mean)
Sensor Mat	62.67	37.33	38.03	24.89
Baseline[6]	100.00	0.00	89.63	10.37
Our	**45.00**	**54.99**	**43.78**	**11.78**

압력 센서 매트 (GT) 값과 비교했을 때, 더 정확한 보행 파라미터 값을 얻을 수 있음을 확인할 수 있다. (표 1 참조).

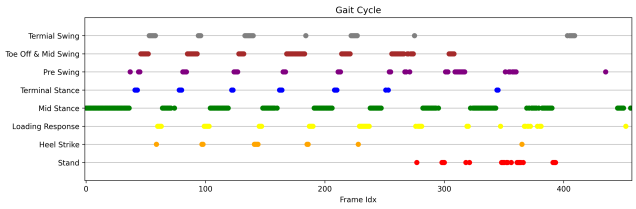


그림 4: 모델의 결과값으로부터 분석한 보행 주기 그래프

4. 결론 및 한계

본 연구에서 2D RGB 영상만으로 보행 주기를 분석하는 딥러닝 기반 프레임워크를 제안하였고, 실험을 통해 그 유의미성을 입증했다. 그러나, 다음과 같은 한계가 존재한다. 보행 주기 분석은 네트워크가 추론한 발과 지면 사이의 접촉 레이블에 의존하는데, 레이블의 품질은 입력 관절 시퀀스의 정확도에 영향을 받는다. 즉, 사용하는 Pose Estimator의 성능에 크게 좌우된다. 따라서, Pose Estimator의 성능이 좋지 않는 경우에 분석된 보행 주기 품질 역시 낮아질 수 있다.

참고문헌

[1] Ri Yu, Hwangpil Park, and Jehee Lee, Human dynamics from monocular video with dynamic camera movements, *ACM Trans. Graph*, 40, 6, Article 208 (December 2021), 14 pages, 2021.

[2] Jiwon Kim, Dongwon Kim, and Ri Yu, Reconstructing Baseball Pitching Motions from Video, *Pacific Graphics Short Papers and Posters*, 2023.

[3] Evans, M., Needham, L., Wade, L. *et al.* Synchronised Video, Motion Capture and Force Plate Dataset for Validating Markerless Human Movement Analysis. *Sci Data,* 11, 1300, 2024.

[4] Lucas Mourot, Ludovic Hoyet, François Le Clerc, Pierre Hellier, UnderPressure: Deep Learning for Foot Contact Detection, Ground Reaction Force Estimation and Footskate Cleanup, *Computer Graphics Forum*, 2022.

[5] CAO, Zhe, et al, OpenPose: Realtime multi-person 2d pose estimation using part affinity fields, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291-7299, 2017.

[6] Rempe, D., Guibas, L., Hertzmann, A., Russell, et al, Contact and Human Dynamics from Monocular Video, *Proceedings of the European Conference on Computer Vision (ECCV), 2020.*

논문 발표

그래픽스응용

양방향 유체 상호작용을 통한 캐릭터 동작 정책의 유체 환경 적용*

남하욱[0], 이윤상
한양대학교 컴퓨터·소프트웨어학과
namhoyog@naver.com, yoonsanglee@hanyang.ac.kr

Applying Character Motion Policies to Fluid Environments via Two-Way Fluid Interaction

Hauk Nam[0], Yoonsang Lee
Dept. of Computer Software Engineering, Hanyang University

요약

본 연구는 파티클 기반 유체 시뮬레이션 환경에서, 캐릭터가 물과 양방향 상호작용을 하며 특정 태스크를 수행하도록 정책을 학습하는 방법을 제안한다. 이를 대표적으로 보여줄 수 있는 사례로 서핑을 목표 태스크로 설정하였으며, 본 논문에서는 그 전 단계로서, 학습된 정책을 활용해 움직이는 파도 위에서 보드 위의 캐릭터가 동작을 수행하는 것을 결과로 보인다.

1. 서론

애니메이션이나 게임과 같은 다양한 분야에서는 사실적인 캐릭터 움직임에 대한 수요가 꾸준히 존재해 왔다. 이러한 수요에 맞추어, 다양한 물리적 환경에서 자연스러운 캐릭터 동작을 생성하는 연구가 지속적으로 이루어지고 있다. 그러나 지금까지의 연구는 주로 지면이나 고체 물체와의 상호작용에 집중되어 있으며, 수영처럼 캐릭터가 물과 상호작용하는 경우를 다룬 연구는 상대적으로 부족하다.

물과 같은 유체를 시뮬레이션하는 기술은 예전부터 연구되어 왔지만, 캐릭터와의 상호작용까지 포함하는 경우는 계산 비용과 시뮬레이션 안정성 문제로 인해 활발히 다뤄지지 않았다. 그 결과, 현재 대부분의 응용 분야에서는 미리 정의된 캐릭터 동작을 사용하며, 물과의 작용은 단방향(one-way)으로만 처리된다. 이로 인해 물과의 상호작용이 제대로 반영되지 못하고, 캐릭터의 동작이 부자연스럽게 표현되는 한계가 있다.

본 연구의 최종 목표는 파티클 기반 유체 시뮬레이션을 활용해 캐릭터가 물과 양방향(two-way)으로 상호작용할 수 있는 환경을 구성하고, 이러한 상호작용을 역동적으로 보여줄 수 있는 서핑 정책을 학습하는 것이다.

* 구두 발표논문, 요약논문 (Extended Abstract)
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회 (논문지)에 제출 준비 중임.
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2023-00222776)과 문화체육관광부 및 한국콘텐츠진흥원의 2025년도 문화체육관광 연구개발사업 지원을 (RS-2024-00399136) 받아 수행되었음.

이에 본 연구는 그 전 단계로서, 지면에서 학습한 정책을 이용해 물 위의 보드에서 파도에 맞춰 넘어지지 않고 정해진 모션을 수행할 수 있음을 보인다. 이를 통해, 지면이나 고체 물체와의 상호작용을 다룬 기존의 모션 학습 정책에서 벗어나, 유체와의 상호작용까지 확장 가능성을 보여준다는 의미를 가진다.

2. 방법

2.1. 유체

유체는 Isaac Lab에서 제공하는 입자 기반 동역학(Particle Based Dynamics, PBD)[1] 방식의 시뮬레이션 기능을 이용해 구현되었다. PBD 방식은 유체를 파티클로 구성하여, 제약(constraint)을 통해 빠른 속도로 현실적인 물리 작용을 만들 수 있다. 유체의 주요 특성인 밀도와 점성은 물의 밀도 1000 kg/m^3와 물의 점성 8.94×10^{-4} $Pa{\cdot}s$로 설정하여 물처럼 시뮬레이션 될 수 있도록 하였다.

2.2. 강화학습

우리는 Masked Mimic[2] 논문의 공식 Github 저장소에 공개되어 있는 사전 학습 모델(pretrained model)을 사용하였다. 이 모델은 AMASS, HumanML3D, SAMP 등 다양한 모션 데이터셋을 종합적으로 사용하여 여러 가지 타입의 고정된 지면에서 학습된 모델이다.

Masked Mimic 방식은 참조 모션을 모방하는 전신 추적 기반의 완전 제약 컨트롤러를 먼저 학습한 뒤, 이를 캐릭터의 일부 정보만으로 전신 동작을 복원할 수 있도록 증류 과정을 통해 부분 제약 컨트롤러로 변환하는 방식이다. 이 일부 정보는 관절 또는 루트의 위치 및 회전 등이 일부 마스크된 참조 모션 형태이다. 이러한 방식으로 학습된 정책은 실행 시 캐릭터에 대한 일부 정보만 참조로 주어도, 다양한 지형에 적응하며 입력된 정보에 맞는 전신 동작을 자연스럽게 생성할 수 있다.

우리는 여러 지면에 적응하도록 학습시킨 pretrained Masked Mimic 모델이 일부 정보만으로도 상황에 적응한 전신 동작을 생성할 수 있기 때문에, 기울기와 높이가 지속적으로 변화하는 유체 기반 보드 위에서도 캐릭터의 중심을 유지하며 안정적인 동작을 생성할 수 있을 것으로 기대하였다.

향후 전이 학습을 통해 서핑에 특화된 동작들을 생성할 수 있을 것이고, high-level 정책을 추가로 학습하여 상황에 맞는 서핑 동작을 할 수 있을 것이라고 기대한다.

2.3. 동적 보드 환경을 위한 정책 입력 재구성

고정된 지면에서 학습된 pretrained 모델을 유체 위에서 움직이는 보드 환경에 적용하기 위해, 관측값 입력과 참조 모션의 기준 위치를 수정하였다. 기존에는 캐릭터 주변의 지형 높이를 관측값으로 사용하였으나, 이를 보드 기준의 높이 정보로 대체하였다. 광선 센서가 파티클은 통과하지만 강체(rigid object)에는 반응하는 특성을 이용하여, 광선이 보드에 닿으면 해당 위치의 보드 높이를, 닿지 않으면 지면 높이를 반환하도록 하였다. 이를 통해 캐릭터는 보드 외곽을 낭떠러지로 인식하게 되며, 보드 위에서 동작하도록 유도한다.
또한, 매 스텝 참조 모션의 루트 위치를 보드 중심 위치에 맞춰 이동시켰다. 이를 통해 캐릭터의 동작이 시작 지점 기준이 아니라, 움직이는 보드를 기준으로 자연스럽게 수행되도록 하였다.

2.4. 파도 생성

현실 세계에서 인공 파도를 생성하는 방식은 다양하다. 일반적으로는 대량의 물을 순간적으로 방출하거나, 펌프 등의 장치를 이용해 특정 지점에서 파형을 만드는 방식이 사용된다. 그러나 Isaac Lab 환경에서 실험한 결과, 이러한 방식으로 생성된 파도는 충분한 거리까지 전달되지 못하고 빠르게 소멸되는 것을 확인하였다.
이에 우리는 서핑에 유의미한 영향을 줄 수 있는, 긴 시간 동안 진행되는 파도를 생성하기 위해 수중에서 강체를 이동시키며 파도를 만드는 방식을 채택하였다. 구체적으로, 그림 1처럼 수면 아래 강체를 인위적으로 움직이면서 유체와 상호작용을 만들고, 이를 통해 그 이동 경로를 따라 파도를 만들어내도록 하였다. 이 방식은 긴 시간 동안 움직이는 파도를 만들 수 있어, 서핑 동작 실험에 보다 적합한 조건을 제공한다.

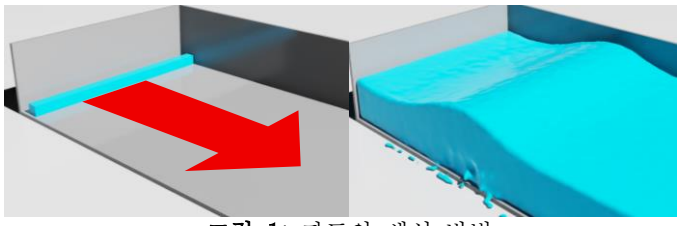


그림 1: 파도의 생성 방법

3. 실험 설계 및 결과

시뮬레이션 환경은 Isaac Lab을 사용하였고, 캐릭터는 Masked Mimic 논문에서 사용된 것과 동일한 neutral SMPL body shape을 기반으로 한 캐릭터를 사용하였다. 보드는 실제 서핑 보드의 물리적 특성을 반영하여 설정하였다. 다만, 현재 Isaac Lab에서는 파티클 기반 유체 시뮬레이션에서는 부력이 실제보다 작게 적용되는 시스템 문제가 있어, 보드 크기를 실제보다 약 2배 크게 설정하였다. 최종적으로 사용된 보드는 길이 5m, 폭 1.5m, 높이 평균 30cm의 롱보드 형태이며, 재질은 EPS 소재의 밀도 15kg/m^3, 마찰계수는 0.8로 설정하였다. 물이 채워진 풀장은 길이 20m, 폭 6m, 높이 3m로 구성하였다. 파도는 수면 아래에서 2m/s의 일정한 속도로 이동하는 강체를 통해, 물체를 따라 낮은 높이의 파도가 생성되도록 하였다. 참조 모션은 AMASS의 CMU 모션 데이터들을 사용하였다.
실험 결과, 그림 2와 같이 파도에 의해 보드가 흔들리고 이동하는 상황에서도 캐릭터는 이에 맞춰서 주어진 참조 모션을 안정적으로 수행하는 것을 확인하였다.

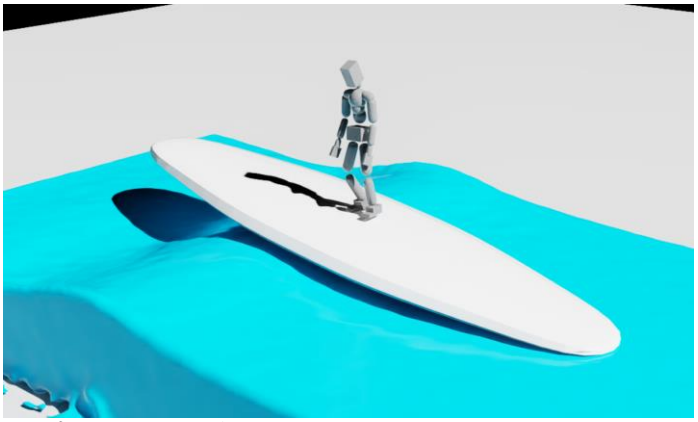


그림 2: 파도 위 움직이는 보드에서 Masked Mimic 모델로 참조 모션(stand motion)을 수행하는 캐릭터

4. 결론

우리는 지면에서 Masked Mimic 방식으로 학습된 모델을 물 위에서 움직이는 보드 환경에 적용하여도 주어진 참조 모션을 환경에 맞추어 안정적으로 생성할 수 있음을 확인하였다.
하지만, Masked Mimic은 구조적으로 입력된 하나의 참조 모션을 기반으로 동작을 생성하므로, 변화하는 환경에 맞춰 능동적으로 동작을 전환하는 데에는 한계가 있다. 이에 향후 연구에서는 변화하는 상황에 맞춰서 task를 수행하기 위해 모션을 선택하는 high-level 정책을 추가로 학습하여, 서핑 동작을 수행할 수 있도록 할 예정이다.

참고문헌

[1] Müller M., Heidelberger B., Hennix M., Ratcliff J. Position Based Dynamics. *Proceedings of the 3rd Workshop in Virtual Reality Interactions and Physical Simulation (VRIPHYS)*, pp. 71–80, 2006.
[2] Tessler C., Guo Y., Nabati O., Chechik G., Peng X.B. MaskedMimic: Unified Physics-Based Character Control Through Masked Motion Inpainting. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2024)*, 43(6), Article 209, 2024.

멀티 에이전트 강화학습 기반 3차원 실내 장면 최적화

조윤식[0,1], 김진모[1,2,*]
한성대학교 일반대학원 정보컴퓨터공학과[0,1], 한성대학교 컴퓨터공학부[2]
yunsik.cho@hansung.ac.kr, jinmo.kim@hansung.ac.kr

3D Indoor Scene Optimization via Multi-Agent Reinforcement Learning

Yunsik Cho[0,1], Jinmo Kim[1,2,*]
Department of Information and Computer Engineering, Graduate School, Hansung University[1],
Division of Computer Engineering, Hansung University[2]

요약

본 연구는 3차원 실내 공간에서의 객체 배치를 자동화하고 최적화하기 위한 멀티 에이전트 강화학습 기반 장면합성 프레임워크를 제안한다. 객체 충돌 회피, 쌍별 객체 기능 유지, 공간 균형 등의 객체 배치 규칙과 사용자 상호작용성을 고려한 요소를 보상 설계에 통합하여 현실적이고 상호작용이 자유로운 배치 유도를 포함한다.

1. 서론

디지털 트윈은 현실 공간을 가상 공간에 복제하여 시뮬레이션을 통해 필요한 정보를 예측하는 기술이다. 이러한 디지털 트윈 기반의 확장현실(XR) 협업 환경에서는 여러 사용자가 동시에 상호작용하므로, 공간 효율성과 동선 등을 고려한 장면 구성이 필수적이다. 장면 최적화 기술은 바로 이러한 물리적, 기능적 특성을 고려하여 자연스럽고 효율적인 가상 장면을 자동으로 최적화하는 데 사용될 수 있다. 기존 장면 최적화 및 생성연구들은 샘플링 기반 최적화[1] 또는 딥러닝 기반[2] 자동 생성 등이 연구되었지만, 이는 연산량이 많거나 데이터셋 의존성이 크기 때문에 유연성에서 한계가 있다. 이처럼 장면 최적화는 명확한 정답 없이 사용자의 주관적인 만족도를 최적화하는 문제이기에, 최근에는 강화학습 기반 장면 최적화 연구도 진행되었다[3]. 이는 에이전트가 주어진 환경(Environment)에서 가능한 행동(Action)을 정의하고, 이를 통해 현재 상태(State)를 관측(Observation)하여 가장 많은 보상(Reward)을 얻을 수 있도록 학습하는 과정을 포함한다. 하지만 기존의 강화학습 기반 장면 최적화 연구들은 단일 에이전트 기반 학습에 그치거나, 옷장의 앞면에 여유공간을 확보하는 단순한 사용자 상호작용성만을 고려할 뿐 다중 사용자 참여 기반 상호작용에 관한 고려까지는 하지 않는 상황이다. 본 연구는 이를 해결하기 위해 다중 객체들을 각 에이전트로 정의하고, 협력적 학습을 통해 공간 최적화를 수행하는 멀티 에이전트 강화학습 기반 프레임워크를 제안하여 배치 타당성과 미적기준은 물론 다중 사용자 상호작용성을 충족시키는 장면 최적화를 제안한다.

* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract)으로서, 본 논문의 연구는 현재 진행 중.
* 교신저자: 김진모(jinmo.kim@hansung.ac.kr)

2. 멀티 에이전트 기반 장면 최적화

본 연구에서는 3차원 실내 장면을 구성하는 객체의 자동 배치를 최적화하기 위해, 멀티 에이전트 강화학습(Multi-Agent Reinforcement Learning) 기반의 장면합성 프레임워크를 제안한다. 각 객체는 독립적인 에이전트로 정의되며, 전체 배치 문제는 마르코프 결정 프로세스(Markov Decision Process)로 정식화된다. 상태 공간은 현재까지 배치된 객체들의 위치, 벽과의 거리, 충돌유무와 같은 의미론적 관계 등을 포함하고, 행동 공간은 다음 단계(step)에 배치할 객체의 위치정보를 의미한다. 에이전트의 학습을 유도하기 위해, 충돌 회피, 기능적 관계 유지(예: 책상-의자), 미적 기준(배치 균형, 밀도 등), 사용자 보행 경로, 여유 공간 같은 다중 사용자 상호작용을 위한 공간확보 등의 요소를 반영한 합성 보상 함수를 설계한다. 각 보상 항목은 중요도에 따라 가중치를 달리하여 종합적인 레이아웃 품질을 반영하도록 구성된다. 정책 학습에는 Proximal Policy Optimization (PPO) 알고리즘을 활용한다. 또한, 객체 간의 협력적 배치를 유도하기 위해, 본 연구에서는 MA-POCA[4] 기반 MPSS(Multi-Agent POsthumous Credit Assignment for Scene Synthesis)를 제안한다. 장면 전체가 완성된 이후 각 객체의 배치 기여도를 기반으로 차등 보상을 분배함으로써, 에이전트 간 협력을 극대화한다.

제안하는 MPSS 알고리즘은 환경 초기화 및 사전에 완전한 장면으로부터 특징을 추출하여 학습에 반영할 수 있도록 하는 Prior 정보 로딩, 에이전트 단위 상태 관측 및 행동 선택, 행동 실행 및 보상 획득, 협력 기여 기반 보상 재할당, 정책 및 가치함수 업데이트, 반복 학습 및 종료 조건 검사의 절차로 진행된다(Algorithm 1). 학습 중 행동 실행 및 보상 획득 과정에서 각 에이전트(배치되는 가구 객체)는 정의된 보상함수를 기반으로 개별 보상을 차등으로 받게 되고, 협력 기여를 기반으로 보상을 재할당하는 과정에서는 전체 장면에서 상호작용이 가능한 공간이 넓을수록 모든 에이전트에게 같은 추가 보상을 부여하여 각 에이전트가 개별적으로 최적의 배치를 찾아가면서도 전체 장면 입장에서 상호작용이 가능한 공간을 확보하려는 방법으로도 학습할 수 있다. 그림 1은 본 연구가 제안하는 MPSS의 개요를 보인다.

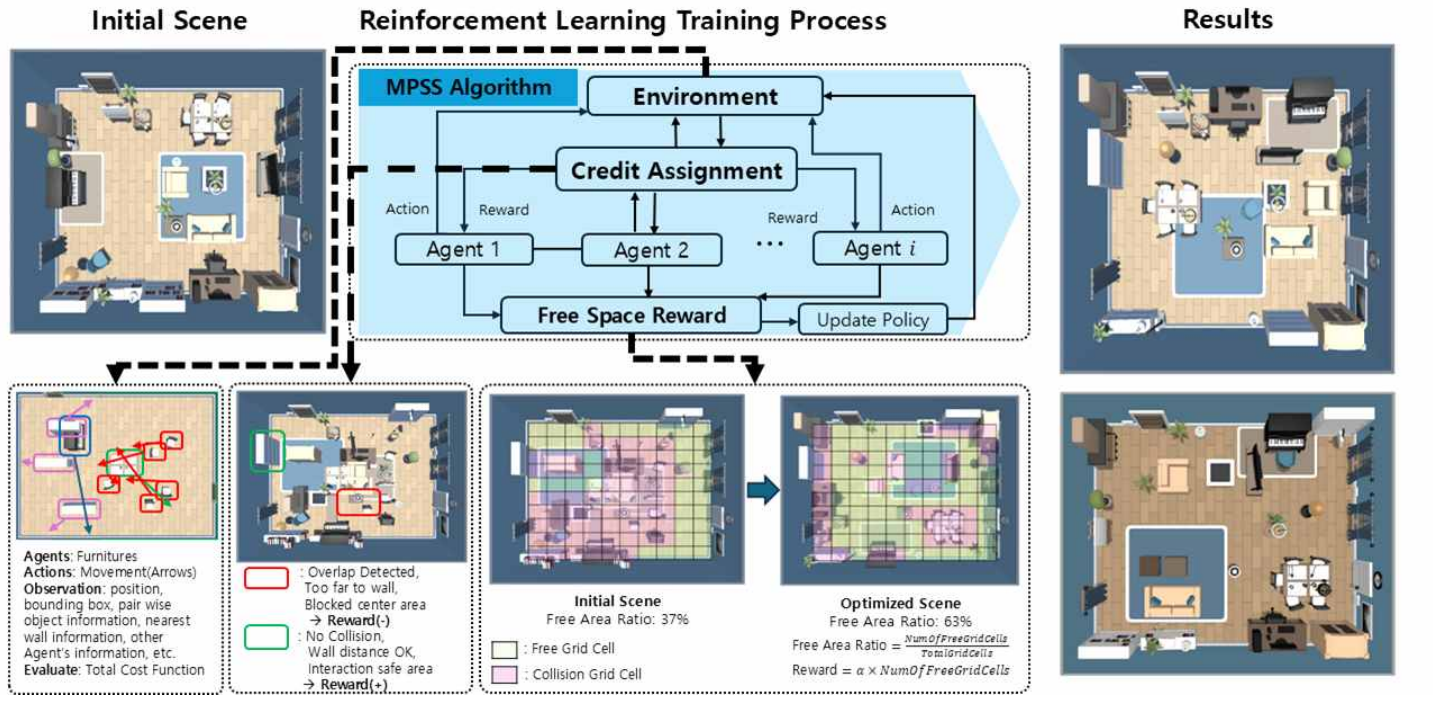


그림 1: Each furniture object is defined as an agent that observes its surroundings, selects movement actions, and receives rewards based on collision avoidance, relational constraints, wall proximity, and free space preservation. The environment incorporates a top-view grid for spatial reasoning and prior layout knowledge to guide optimization.

알고리즘 1. Multi-agent Posthumous credit assignment for Scene Synthesis (MPSS)

1. **procedure** Object Optimization
2. Initialize scene and agent list
3. Load prior layout information
4. **for each** episode **do**
5. Reset environment
6. **while** not all agents placed **do**
7. **for each** agent **do**
8. Observe state s_t
9. Select action $a_t \sim \pi_\theta(a_t \mid s_t)$
10. Execute action and update position
11. Compute reward r_t: includes collision, relation, wall distance, placement, etc.
12. Evaluate contribution weight for each agent
13. Compute grid-based free space reward
14. Compute advantage A_t using MA-POCA
15. Update π_θ, V_ϕ(critic) using PPO, TD loss
16. **end for**
17. **end while**
18. **end for**
19. **end procedure**

3. 결론

본 연구에서는 다수의 객체가 상호 의존적인 관계를 맺으면서 다중 사용자의 상호작용을 고려한 3차원 협업 공간을 최적화하기 위해, 멀티 에이전트 강화학습 기반의 장면 최적화 프레임워크 MPSS를 제안하였다. 이를 위해 각 객체를 지능형 에이전트로 모델링하고, 이들이 순차적인 협상 프로토콜을 통해 상호작용하며 최적의 배치를 스스로 학습하도록 설계 하였다. 향후 실내 가구 배치를 넘어, 스마트 팩토리의 설비 최적화와 같은 객체 간의 유기적인 관계가 중요한 다양한 디지털 트윈 분야로 확장될 수 있을 것으로 기대한다.

참고문헌

[1] Yu, L. F., Yeung, S. K., Tang, C. K., Terzopoulos, D., Chan, T. F., & Osher, S. J., "Make it home: automatic optimization of furniture arrangement," *ACM Transactions on Graphics*, 30,(4), 86, 2011.

[2] Paschalidou, D., Kar, A., Shugrina, M., Kreis, K., Geiger, A., & Fidler, S., "Atiss: Autoregressive transformers for indoor scene synthesis," *Advances in Neural Information Processing Systems*, 34, 12013-12026, 2021

[3] Sun, J. M., Yang, J., Mo, K., Lai, Y. K., Guibas, L., & Gao, L., "Haisor: Human-aware indoor scene optimization via deep reinforcement learning," *ACM Transactions on Graphics*, 43(2), 1-17, 2024.

[4] Cohen, A., Ervin, T., Vincent-Pierre, B., Ruo-Ping, D., Hunter, H., Marwan, M., Alexander, Z., Sujoy, G., "On the use and misuse of absorbing states in multi-agent reinforcement learning," *RL in Games Workshop AAAI 2022,* https://arxiv.org/abs/2111.05992.

논문 발표

렌더링/이미지/비디오

포트레이트토커: 텍스트 프롬프트 기반 음성 구동 3D 말하는 얼굴 생성*

DU XIAN[0,1], 유리[1,2,*]
아주대학교 인공지능학과[1], 아주대학교 소프트웨어학과[2]
duxian@ajou.ac.kr, riyu@ajou.ac.kr

PortraitTalker: Speech-Driven 3D Talking Head from Text Prompt

XIAN DU[0,1], Ri Yu[1,2,*]
Dept. of Artificial Intelligence, Ajou University[1], Dept. of Software and Computer Engineering, Ajou University[2]

Abstract

The reliance on reference images or 3D models presents a fundamental limitation for customizable digital avatar creation. We propose PortraitTalker, an end-to-end framework that generates photorealistic 3D avatars directly from text prompts and speech inputs without requiring manual rigging. Our system integrates a diffusion model with score distillation sampling for texture generation and a transformer-based audio encoder to drive FLAME-based facial animation, PortraitTalker achieves state-of-the-art performance on the HDTF dataset, improving lip synchronization (LSE-C: 7.230 , LSE-D: 7.712) and visual quality (FID: 21.997). This work advances automated avatar creation by removing conventional input constraints, enabling scalable applications in AR/VR and intelligent virtual agents.

1. Introduction

Digital avatars are playing an increasingly important role in immersive applications. However, current generation techniques are often constrained by their reliance on input images or predefined 3D templates. Although significant advancements have been made in text-to-3D synthesis [1] and speech-driven animation [2] independently, integrating both into a cohesive pipeline remains challenging, particularly in preserving temporal consistency and visual fidelity.

PortraitTalker addresses these challenges through a unified architecture composed of 3 key components:

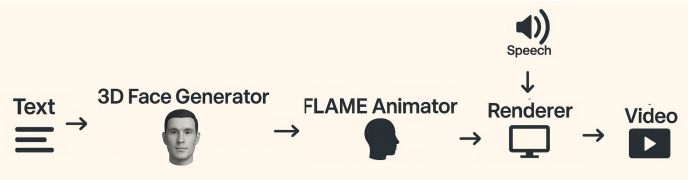


Figure 1: Pipeline of the PortraitTalker

(1) Text-to-3D Synthesis: SDS-optimized diffusion enables high-quality 3D texture generation from textual descriptions;
(2) Speech-Driven Animation: A transformer-based audio processor extracts FLAME-compatible parameters for expressive facial motion;
(3) Differentiable Rendering: A real-time renderer ensures temporal coherence and physical plausibility in the final output.

2. Methodology

2.1. Text-to-3D Synthesis

Our pipeline initiates avatar creation via SDS-optimized diffusion, which distills gradients from a pretrained text-to-image model. This generates a tri-grid representation that jointly encodes geometry and texture. Orthogonal feature planes facilitate efficient synthesis of animation-ready models with high visual fidelity.

2.2. Speech-Driven Animation

A transformer-based audio encoder is employed to predict frame-wise FLAME parameters [3] directly from raw speech input. These parameters capture both expression dynamics and head pose variations. The result is accurate, temporally aligned lip-sync and expressive motion patterns that reflect the speech signal.

* 구두발표논문
* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2025-RS-2023-00255968)

2.3. Differentiable Rendering

Our rendering pipeline employs a differentiable renderer to synthesize video frames. It composites the FLAME-based geometry with hierarchically structured tri-grid textures. The renderer inherently enforces spatiotemporal coherence and physically accurate shading, eliminating the need for post-processing stages.

3. Experiments

3.1. Qualitative Comparison

Our method demonstrates statistically significant improvements across all evaluation metrics on the HDTF dataset [6]. As quantified in Table 1, PortraitTalker achieves a Lip Sync Error Confidence (LSE-C) score of 7.230, representing relative improvements of 43.2% and 48.4% over MakeItTalk [4] (5.051) and Wang et al. [5] (4.872) respectively. Concurrently, we reduce the Lip Sync Error Distance (LSE-D) by 22.9% (7.712 vs. 9.999/9.995), indicating superior temporal alignment accuracy. In terms of visual quality, our approach establishes a new state-of-the-art Frechet Inception Distance (FID) of 21.997, outperforming both baselines by significant margins.

Method	Lip Synchronization		Video Quality
	LSE - C↑	LSE - D↓	FID↓
MakeItTalk [4]	5.051	9.999	28.183
Wang et al. [5]	4.872	9.995	22.372
Ours	7.230	7.712	21.997

Table 1. Comparison with methods on HDTF [6] dataset

3.2. User Study

We conducted a comprehensive user evaluation with 20 participants assessing 50 generated video samples. As shown in Table 2, our method achieved dominant preference scores across four key perceptual metrics: lip-sync accuracy (68.13% preference), motion diversity (76.89%), video sharpness (74.06%), and overall naturalness (74.76%). Notably, 38% of participants explicitly identified our system as superior specifically for lip-sync quality.

Method	Lip Sync.	Motion Diversity	Video Sharpness	Overall Naturalness
MakeItTalk[4]	9.86%	7.04%	6.72%	9.41%
Wang et al.[5]	22.01%	16.07%	19.22%	15.83%
Ours	68.13%	76.89%	74.06%	74.76%

Table 2: User Study

3.3. Results

Figure 2 showcases five representative keyframes synthesized from the prompt “A casually dressed young adult European male”. The outputs show temporally coherent facial expressions, phoneme-level lip-sync, and consistent identity preservation throughout the animation sequence.

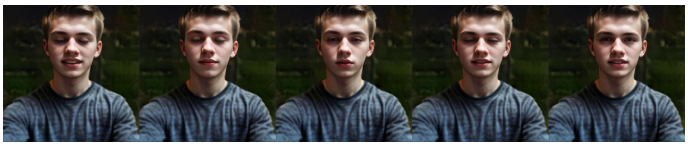


Figure 2: Result of the PortraitTalker.

4. Conclusion

We present PortraitTalker, a novel framework for generating photorealistic, speech-driven 3D avatars from text prompts. By eliminating the need for reference images and manual rigging, Our system enables the creation of high-quality, scalable avatars. Experimental results on both objective metrics and user studies demonstrate superior performance in lip-sync accuracy and video realism. Future work includes enhancing emotional expressiveness, incorporating neural shading for greater realism, and optimizing the model for lightweight real-time deployment.

Reference

[1] Wu Y, Xu H, Tang X, et al. Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior[J]. ACM Transactions on Graphics (TOG), 2024, 43(4): 1-12.
[2] Zhang W, Cun X, Wang X, et al. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 8652-8661.
[3] Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4D scans[J]. ACM Trans. Graph., 2017, 36(6): 194:1-194:17.
[4] Zhou Y, Han X, Shechtman E, et al. Makelttalk: speaker-aware talking-head animation[J]. ACM Transactions On Graphics (TOG), 2020, 39(6): 1-15.
[5] Wang S, Li L, Ding Y, et al. One-shot talking face generation from single-speaker audio-visual correlation learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 2531-2539.
[6] Zhang Z, Li L, Ding Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3661-3670.

Octree를 이용한 광선추적기반 3D Gaussian LOD 제어*

박지영, 김영준
이화여자대학교 컴퓨터공학과
{jiyoung_06, kimy}@ewha.ac.kr

Level-of-Detail Control for 3D Gaussian Ray Tracing using an Octree

Jiyoung Park, Young J. Kim
Dept. of Computer Science and Engineering, Ewha Womans University

요약

본 연구에서는 3D Gaussian 렌더링을 위한 ray tracing 기반 실시간 렌더링 기법에 계층적 표현 방식을 도입하기 위하여 octree 구조를 이용하는 방법을 제안한다. 즉, Gaussian에 octree level 정보를 부여하여 ray tracing에 level-of-detail 표현을 구현하였다. 또한 octree 기반의 가시성 마스크 계산을 OptiX 기반 파이프라인에 적용하여 렌더링 성능을 향상했다. 제안하는 방법은 기존의 방법과 유사한 품질을 유지하면서도 렌더링에 필요한 Gaussian 수는 최대 80% 감소하며, 렌더링 속도 또한 최대 120% 향상된 것을 확인하였다.

1. 서론

최근 들어 Gaussian splatting 기반의 3D 장면 재구성 [1]이 높은 렌더링 품질과 실시간 처리 속도를 동시에 달성하는 방법으로 주목받고 있다. Moenne-Loccoz et al. [2]가 제안한 3D Gaussian ray tracing(3DGRT)은 Gaussian splatting을 ray tracing 방식으로 렌더링한 것으로, 다양한 시점이나 왜곡된 카메라에 대응할 수 있는 ray tracing의 보다 유연한 렌더링 능력을 유지하면서도 실시간 성능을 달성할 수 있음을 보였다. 그러나 Gaussian의 수가 많아질수록 렌더링에 필요한 연산과 메모리 사용량이 증가하여 실시간성에 제약이 발생하는 문제가 있다.

한편, Octree-GS [3]는 Gaussian을 계층적으로 관리하는 구조를 통해 level-of-detail(LOD)을 효율적으로 제어하며, 복잡한 장면을 비교적 적은 수의 Gaussian으로 표현할 수 있다. 하지만, 이 방식은 rasterization 기반으로 설계되어 있어, ray tracing 기법과는 직접적으로 호환되지 않는다.

본 연구에서는 3DGRT의 실시간 ray tracing 기반 렌더링 구조에 계층적 Gaussian 표현법을 통합하였다. 이를 위해 octree의 voxel에 대응하는 위치와 level 정보를 3DGRT 파이프라인에 도입하여, 계층적 Gaussian 표현과 LOD 제어가 ray tracing 환경에서도 가능하도록 구현하였다. 이 논문에서 제안하는 Octree-3DGRT에서는 장면을 표현하기 위한 Gaussian의 수를 실외 장면의 경우 최대 80% 줄이고, 렌더링 속도를 최대 120% 높이면서도 실시간 렌더링 품질을 유지할 수 있었다.

* 학부생 주저자 논문, 구두 발표 초록 논문 (extended abstract).
* 본 연구는 과학기술정보통신부의 재원으로 정보통신기획평가원 ITRC의 지원(IITP-2025-RS-2020-II201460)과 한국연구재단 (2022R1A2B5B03001385)의 지원을 받아 수행되었음.

2. LOD기반의 3DGRT

2.1. 계층적 Gaussian 표현

Octree-GS에서는 Gaussian의 위치를 octree의 voxel로 나타내어, 각 Gaussian에는 LOD 제어를 위해 voxel의 level 정보가 포함된다. 또한 Gaussian의 분포는 differential Gaussian 렌더링 과정 중 주기적으로 수행되는 densification 단계에서 동적으로 확장된다. 기본적으로 새로운 Gaussian은 선택된 Gaussian과 octree 상에서 같은 level의 voxel에 생성되지만, 더욱 세밀한 표현이 필요한 경우 하위 level의 voxel에서 생성된다.

이때 densification을 위한 Gaussian의 선택 기준으로는 위치 gradient가 사용된다. Differential 렌더링이 ray tracing으로 이루어지기 때문에 Octree-GS의 스크린 공간 좌표계가 아닌 3DGRT의 3차원 월드 좌표계의 위치 gradient를 사용한다. 또한, 위치 gradient에 대한 임곗값은 level에 따라 단계적으로 조정된다. 이렇게 결정된 level은 렌더링 단계에서 시점에 따른 LOD를 제어하는 데 활용된다.

2.2. LOD 기반 가시성 마스크 생성

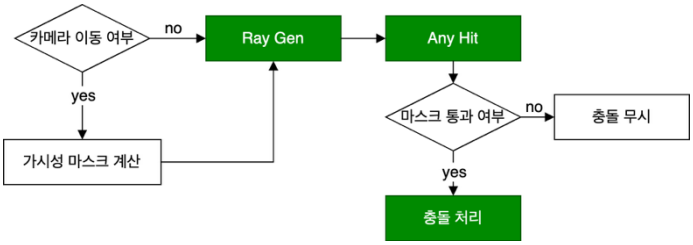


그림 1. 3DGRT 렌더링 파이프라인 내 가시성 마스크 삽입

3DGRT는 Gaussian의 수가 많아질 경우, 실제로 해당 Gaussian이 렌더링 결과에 크게 영향을 주지 않더라도 alpha 값이 충분히 누적될 때까지 광선을 따라 선적분을 수행하여 연산량이 증가하는 문제가 있다. 이를 보완하기 위해 본 연구에서는 3DGRT의 ray tracing 파이프라인에 가시성 마스크 계산 단계를 도입한다.

그림 1과 같이, 카메라가 이동할 때마다 각 Gaussian에 대해 카메라와의 거리 및 해당 Gaussian의 level 정보를 바탕으로 가시성을 판단하고, 이를 통해 전체 Gaussian에 대한 마스크를 생성한다. 생성된 마스크는 이후 광선 교차 확인 시점인 Any Hit 쉐이더 연산에서 렌더링 대상 여부를 결정하는 기준으로 사용된다. 이를 통해, 불필요한 Gaussian을 파이프라인 앞단에서 제외함으로써 전체 연산량을 효과적으로 줄일 수 있다.

3. 실험 결과

	bicycle	bonsai	counter	garden	kitchen	room	stump
PSNR	**24.766**	**31.545**	28.369	26.593	29.503	30.101	26.249
SSIM	**0.742**	**0.936**	**0.899**	0.842	**0.913**	0.904	**0.765**
LPIPS	**0.255**	**0.246**	**0.257**	**0.145**	**0.168**	**0.297**	**0.251**
FPS	45.662	51.813	45.809	49.456	29.869	54.555	47.687
#G(K)	5,744	1,558	**1,284**	4,195	**1,541**	1,297	6,770

표 1. 3DGRT의 실험 결과

	bicycle	bonsai	counter	garden	kitchen	room	stump
PSNR	24.560	30.198	**28.457**	**26.802**	**29.525**	**30.559**	**26.264**
SSIM	0.733	0.916	0.896	**0.848**	0.910	**0.910**	0.758
LPIPS	0.288	0.289	0.266	0.146	0.177	0.299	0.285
FPS	**87.719**	**54.675**	**52.383**	**51.948**	**33.750**	**77.399**	**105.263**
#G(K)	**2,011**	**1,556**	1,487	**4,189**	2,671	**1,109**	**1,335**

표 2. Octree-3DGRT의 실험 결과

	bicycle	bonsai	counter	garden	kitchen	room	stump
FPS	92.11	5.52	14.35	5.04	12.99	41.87	120.74
#G	64.99	0.15	-15.74	0.14	-73.33	14.50	80.28

표 3. FPS 증가율(%), #G 감소율(%)

제안하는 연구의 실험은 NVIDIA RTX 6000 Ada GPU를 사용한 환경에서 수행되었으며, PyTorch 기반으로 구현되었다. Gaussian 렌더링 및 ray tracing 연산에는 NVIDIA OptiX 엔진을 활용하였고, 공개 데이터셋인 Mip-NeRF360의 7개 장면을 사용하여 성능을 비교 검증하였다. 품질 평가지표로 PSNR, SSIM, LPIPS를 사용하였고, 성능 평가지표로는 렌더링 FPS와 사용된 Gaussian의 수(#G(K))를 사용하였다.

실험 결과, 표 2에서와 같이 PSNR, SSIM, LPIPS 등 품질 지표는 기존 3DGRT(표 1)와 유사한 수준을 유지했다. 한편, 표 3에서 볼 수 있듯 FPS 증가율과 Gaussian 수의 감소율은 장면에 따라 편차를 보였다.

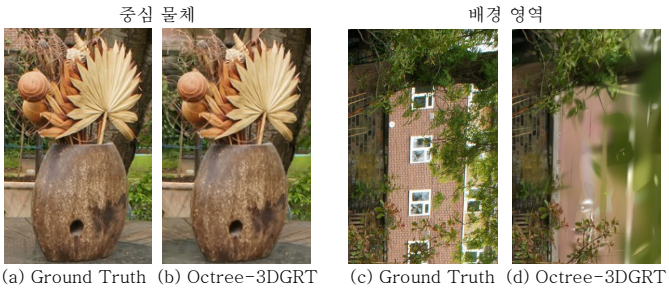


그림 2. 배경 영역과 중심 물체 비교

예를 들어, 실외 장면인 bicycle, stump는 배경이 복잡하게 구성되어 Gaussian의 수가 약 65~80% 감소하였고, 이에 따라 FPS도 92~120%가량 증가하였다. 반면, 중심 물체 위주로 구성된 실내 장면인 counter, kitchen에서는 Gaussian의 수가 오히려 증가하였으나, LOD 기반 마스킹 효과로 인해 FPS는 약 13% 증가하는 결과를 보였다.

또한 실외 장면의 경우, 그림 2(a, b)에서 보이듯 중심에 위치한 물체는 높은 재구성 품질을 보였지만, 그림 2(c, d)와 같이 학습 데이터에 충분히 포함되지 않은 배경 영역은 흐릿하게 표현되는 문제가 발생하였다. 이는 densification 과정에서의 가시성 필터링 조건이 영향을 준 것으로 보이며, 향후 이에 대한 개선이 필요하다.

4. 결론

본 연구에서는 ray tracing 기반 3D Gaussian 렌더링 파이프라인에 계층적 Gaussian 표현 방식을 통합하는 방법을 제안하였다. 그 결과, 렌더링 품질 지표는 기존과 유사한 수준을 유지하였으며, 이에 따라 평균적으로 프레임당 렌더링 시간 또한 감소하였다.

향후에는 현재 LOD를 위한 octree 구조를 확장하여, 시점에 따라 필요한 Gaussian만을 선택적으로 GPU에 적재하는 out-of-core 방식으로 발전시킴으로써, 대규모 장면에 대한 실시간 렌더링 효율을 더욱 향상할 방법을 연구할 계획이다.

참고문헌

[1] Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G, 3D Gaussian Splatting for Real-Time Radiance Field Rendering, *ACM Trans. Graph.*, 42(4):139-1, July 2023.
[2] Moenne-Loccoz, N., Mirzaei, A., Perel, O., de Lutio, R., Martinez Esturo, J., State, G., Fidler, S., Sharp, N., and Gojcic, Z, 3D Gaussian Ray Tracing: Fast Tracing of Particle Scenes, *ACM Trans. Graph.*, 43(6):1-19, November 2024.
[3] Ren, K., Jiang, L., Lu, T., Yu, M., Xu, L., Ni, Z., and Dai, B., Octree-GS: Towards Consistent Real-Time Rendering with LOD-Structured 3D Gaussians, *IEEE Trans. Pattern Anal. Mach. Intell.*, doi:10.1109/TPAMI.2025.3568201, May 2025.

Diffusion 선행 지식을 활용한 시공간 일관적 비디오 초해상도 기법*

한장혁[†,1], 심규진[0,†,1], 김건웅[1], 이현승[2], 최규하[2], 한영석[2], 조성현[1]
[1]포항공과대학교, [2]삼성전자 VD 사업부
{hjh9902, sgj0402, k2woong92}@postech.ac.kr, {hyuns.lee, kyuha75.choi, yseok.han}@samsung.com, s.cho@postech.ac.kr

DC-VSR: Spatially and Temporally Consistent Video Super-Resolution with Video Diffusion Prior

Janghyeok Han[†,1], Gyujin Sim[0,†,1], Geonung Kim[1],
Hyun-seung Lee[2], Kyuha Choi[2], Youngseok Han[2], Sunghyun Cho[1]
[1]POSTECH, [2]Visual Display Business, Samsung Electronics

Abstract

Video Super-Resolution (VSR) aims to restore high-resolution (HR) videos from low-resolution (LR) counterparts. While recent diffusion-based VSR methods exhibit remarkable performance, their tile-based processing and inherent randomness often cause severe inconsistencies across the spatio-temporal domain. To address this, we introduce DC-VSR, a novel VSR method for generating spatially and temporally consistent videos with high-fidelity textures. Specifically, to ensure spatio-temporal consistency, DC-VSR proposes Spatial Attention Propagation (SAP) and Temporal Attention Propagation (TAP), which propagate information across spatio-temporal tiles using a self-attention mechanism. To further enhance high-frequency details, we present Detail-Suppression Self-Attention Guidance (DSSAG), a novel diffusion guidance scheme. DC-VSR achieves high-quality, spatio-temporally consistent VSR, outperforming prior methods.

1. Introduction

Video Super-Resolution (VSR) focuses on restoring high-resolution (HR) videos from low-resolution (LR) counterparts. To synthesize sharp details, recent approaches[1,2] utilize the powerful generative prior of diffusion models. To reduce excessive memory usage, previous methods adopt an independent tile-based processing strategy; however, this leads to spatio-temporal inconsistency due to the lack of shared contextual information across tiles. Moreover, the inherent randomness of the diffusion process further degrades temporal consistency, causing flickering artifacts.

In this paper, we introduce DC-VSR, a novel VSR method to achieve high-resolution video generation with improved spatial and temporal consistency. To this end, we adopt a video diffusion prior for the first time, which better captures temporal dynamics across frames than image diffusion models. We also propose Spatial Attention Propagation (SAP) and Temporal Attention Propagation (TAP) that propagate information across spatio-temporal tiles using a self-attention mechanism to achieve high spatio-temporal consistency. Finally, to further enhance high-frequency details, we propose Detail-Suppression Self-Attention Guidance (DSSAG), steering the model to focus more on high-frequency components.

2. DC-VSR

Fig. 1(a) illustrates the overall pipeline of DC-VSR. DC-VSR starts by bicubic upsampling the LR video and encoding it into a latent representation l. The HR latent x_T is initialized with random noise and denoising proceeds iteratively: at each step t, x_t and l are concatenated, tiled, denoised, and merged. For computational efficiency, SAP and TAP are applied alternately in the denoising step. At the end of each denoising step, DSSAG is applied to make sharp details.

† 공동 제1저자
* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 SIGGRAPH 2025 Conference Paper에 게재 확정되었음.
* 본 연구는 삼성전자의 지원을 받아 수행되었음. 본 연구는 또한 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었음 (RS-2019-II191906, 인공지능대학원지원(포항공과대학교), No.2021-0-02068, 인공지능 혁신 허브 연구 개발).

Finally, denoised x_0 is decoded to yield the HR video.

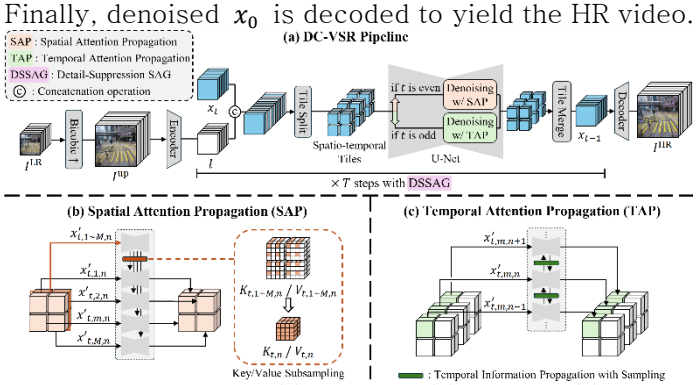


Figure 1: Pipeline of (a) DC-VSR, (b) SAP and (c) TAP.

2.1. Spatial Attention Propagation

Spatial Attention Propagation (SAP) strengthens consistency across spatial tiles by allowing each tile to access information from other spatial tiles using self-attention mechanism (Fig. 1(b)). Instead of calculating time-consuming full attention, SAP uses subsampled key/value features from other spatial tiles. Subsampled key/value features are combined with each tile's own key/value features at self-attention layers to provide global spatial context.

2.2. Temporal Attention Propagation

Temporal Attention Propagation (TAP) enhances temporal consistency by enabling each temporal tile to utilize neighboring temporal tile information (Fig. 1(c)). TAP subsamples the most informative key/value features from the previous and succeeding temporal tiles based on the key feature variance and incorporates subsampled features into the attention mechanism of the current temporal tile. In this manner, the model can leverage bidirectional temporal context to produce temporally consistent results.

2.3. Detail-Suppression Self-Attention Guidance

Detail-Suppression Self-Attention Guidance (DSSAG) is a novel diffusion guidance method designed to enhance high-frequency details in VSR. By suppressing details in the unconditional noise within the CFG[5], the model is directed to focus more on high-frequency components, promoting sharper outputs. Details are suppressed in the self-attention layers within detail-suppression self-attention, which is defined by the following equation:

$$\text{Self-Attention}(Q,K,V,\gamma) = \text{softmax}\left(\frac{QK^T}{\max(\gamma^2 qk, 1)\sqrt{d}}\right)V,$$

where q and k denote the largest absolute values in Q and K, and d is the attention feature dimension. The detail suppression strength, governed by γ, decays over time and can be tailored to target video for granular control. DSSAG within CFG is defined as:

$$\varepsilon_{CFG\&DSSAG}(x_t) = \varepsilon'_\theta(x_t) + (1+s)(\varepsilon_\theta(x_t, c) - \varepsilon'_\theta(x_t)),$$

where ε_θ is a pre-trained U-Net and ε'_θ is a pre-trained U-Net equipped with detail suppressed self-attention. s denotes the guidance scale. DSSAG can be seamlessly incorporated into the CFG without any additional U-Net inference overhead.

3. Result

DC-VSR surpasses prior state-of-the-art methods[1,2,4] by delivering higher perceptual quality and stronger spatio-temporal consistency. This superiority is clearly demonstrated in Fig. 2. The proposed SAP, TAP and DSSAG methods effectively improve VSR results as shown in Tab. 1.

Figure 2: Comparison with other state-of-the-art methods.

SAP	TAP	DSSAG	MUSIQ↑	DOVER↑
-	-	-	67.51	66.07
○	-	-	67.52	66.10
-	○	-	67.51	66.19
○	○	○	69.22	70.41

Table 1: Ablation study on SAP, TAP and DSSAG.

References

[1] Zhou, Shangchen, et al. "Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[2] Yang, Xi, et al. "Motion-guided latent diffusion for temporally consistent real-world video super-resolution." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
[3] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." arXiv preprint arXiv:2207.12598 (2022).
[4] Zhang, Yuehan, and Angela Yao. "Realviformer: Investigating attention for real-world video super-resolution." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

논문 발표

시뮬레이션/모델링

GarmentoPIA: 지능형 에이전트를 활용한 의복 패턴 모델 생성 시스템*

염민기[0], 신림수, 이성희
한국과학기술원 문화기술대학원
{mingiyeom, pogback, sunghee.lee}@kaist.ac.kr

GarmentoPIA: Generating Garment Pattern Models using Intelligent Agents

Mingi Yeom[0], Rimsoo Shin, Sung-Hee Lee
Graduate School of Culture Technology, KAIST

요약

본 연구는 대형 언어 모델(LLM) 지능형 에이전트 기반 파라 메트릭 의복 패턴 모델을 생성하는 GarmentoPIA를 제안한다. 사용자는 원하는 의복 제도 문헌을 선택하고, GarmentoPIA를 이용하여 해당 문헌에 명시된 모든 재봉 파라미터를 반영한 의복 패턴 모델을 생성한다. 다양한 문헌에 대한 범용성과 안정성을 확보하기 위해, 본 시스템은 의복 질의 생성 모듈(Garment Query Generation Module)을 이용하여 시스템의 프롬프트를 자체적으로 개선한다. 더불어, 패턴 도식의 각 구성 요소에 적합한 함수들로 이루어진 새로운 의복 도메인 특화 언어(DSL)를 도입한다. DSL은 GarmentoPIA가 패턴 모델을 손쉽게 생성할 수 있도록 지원하며, 생성된 의복 패턴 모델이 LLM 없이 동작할 수 있도록 한다.

1. 서론

의복은 2차원 제도 패턴을 바탕으로 제작되며, 이러한 패턴은 봉제 후 3차원 형태로 완성된다. 최근 고품질 의복 모델링에 대한 관심이 높아지면서, 의복 패턴을 생성하는 다양한 파라 메트릭 의복 패턴 모델[1, 2]이 개발됐으나, 대부분 수작업으로 설계되어 스타일적·기능적 다양성을 충분히 반영하지 못하는 한계가 있다. 본 연구에서는 LLM을 기반으로 기존 패턴 제작 문헌에서 제도 지식을 추출·구조화하여, 실제 재봉에 사용되는 패턴과 유사한 모델을 자동으로 생성하는 GarmentoPIA 시스템을 제안한다 (그림 1 참조). GarmentoPIA는 문헌 기반 패턴 모델링을 통해 다양한 스타일과 복잡성을 갖는 의복 패턴 모델을 효율적으로 생성할 수 있으며, 생성된 모델은 LLM에 독립적인 프로그램 형태로 제공되어 활용성과 접근성이 뛰어나다.

GarmentoPIA는 SQL 기반 에이전트와 검색 증강 생성(RAG) 기법을 활용한 의복 질의 생성 모듈, DSL 검증기(DSL Proofreader), 최종 실행할 수 있는 패턴 모델로 조직하는 후처리기로 구성된다. 각 요소는 패턴 논리를 표현하는 DSL을 사용한다. 이를 통해 사용자는 다양한 패턴 제도 문헌을 기반으로 맞춤형 패턴 모델을 생성할 수 있으며, 고품질 의복 모델링에 필요한 기술적·자원적 장벽을 크게 낮출 수 있다. 본 연구의 주요 기여는 최초로 의복 문헌 기반의 파라 메트릭 패턴 모델 생성기를 제안한 점, 자체 프롬프트 개선을 통한 질의 생성 모듈의 성능 향상, 문헌에 명시된 모든 제도 및 신체 치수를 지원하는 패턴 모델을 제공하는 점이다.

그림 1: 의복 제도 문헌 기반 의복 패턴 생성

2. 방법론

2.1. 전체 구조

제시하는 방법론은 데이터 전처리와 GarmentoPIA로 구성된다 (그림 2 참조). 데이터 전처리에서는 의복 패턴 제도 데이터를 획득하기 위해 Modern Pattern Design[3]을 참고하여 패턴 또는 패널 단위로 구조화한다. GarmentoPIA는 구조화된 제도 데이터를 이용하여 의복 패턴 모델을 생성한다. 생성한 의복 패턴 모델은 사용자의 입력을 바탕으로 의복 패턴을 생성한다.

2.2. 도메인 특화 언어(DSL) 함수 선언

의복 제도 작업을 명확하고 모듈화된 방식으로 공식화하기 위해 DSL을 제안한다. 이 DSL은 전통적인 패턴 제작 도구(자, 컴퍼스 등)에서 영감을 받아, 직선, 원,

* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회(논문지)에 제출중
* 본 연구는 한국전파진흥원 메타버스랩 지원사업으로 수행되었음.
* 이 논문은 2025년 '2025년 문화체육부의 재원으로 2025 지역특화콘텐츠개발지원사업 (대전)' 지원을 받아 수행된 연구임.
* 이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "패션상품 및 착용 영상"을 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

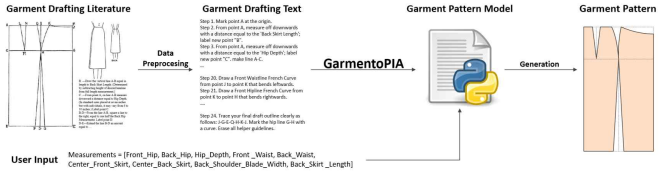


그림 2: GarmentoPIA를 이용한 의복 패턴 생성 절차

프렌치 커브 등 다양한 형태의 제도 연산을 함수적 추상화로 구현한다. 각 DSL 함수는 점, 선분, 곡선 등 제도 구성 요소를 생성하며, 함수 선언 집합(D_d)와 연산 우선순위 집합(D_o)로 구성된다. D_d에는 다양한 도형 생성 및 교차점 탐색 함수가 포함되며, 모든 매개변수에 타입 힌트를 제공하여 구문 및 의미 검증이 가능하다. D_o는 연산 복잡도에 따라 실행 순서를 지정한다. 프렌치 커브는 베지어 곡선으로 체계화하여 문헌 기반 파라미터로 정의한다. 또한, LLM 기반 자동화 과정에서 발생할 수 있는 비결정적 명명 문제를 해결하기 위해, GarmentoPIA는 구성 요소와 함수에 일관된 명명 규칙을 사용한다.

2.3. 의복 질의 생성 모듈

GarmentoPIA의 전체 구조는 의복 질의 생성 모듈을 중심으로 패턴 구조 DB(Pattern Component DB)를 구축하는 방식으로 설계되어 있다. 이 모듈은 Draft2DSL 번역기(Draft2DSL Translator), 의복 구조 질의 생성기(Component Query Generator), 의복 구조 질의 검토기(Component Query Reviewer)의 세 가지 전문 에이전트로 구성된다. 각 제도 단계에서 Draft2DSL 번역기가 코드 스니펫을 생성하고, 의복 구조 질의 생성기가 패턴 구성 요소(점, 선, 상수 등)를 대기 리스트 테이블에 추가한다. 이후 의복 구조 질의 검토기가 해당 데이터를 평가하여, 대기 리스트가 비어 있으면 빠진 요소를 식별하고, 데이터가 채워져 있으면 정확성을 검증한다. 문제가 발견되면 검토기는 보조 규칙과 예시를 추가하여 프롬프트를 개선하고, 이 과정을 반복한다. 검증이 완료된 데이터는 최종 구조 테이블로 이동하며, DSL 검증기(DSL Proofreader)가 일관성 검사를 수행한 후, 후처리기가 최종 테이블의 데이터를 바탕으로 실행할 수 있는 의복 패턴 모델을 생성한다.

3. 결과

본 연구에서 생성된 패턴 구조 DB의 무결성을 검증하기 위해 문법이 실행할 수 있는 Python 코드로 얼마나 정확히 표현되는지를 평가하였고, 의복 모델 내 구성 요소에 할당된 값이 의복 패턴 제도 텍스트에 기술된 정보를 얼마나 충실히 반영하는지를 검토한다. 실험에서는 SMPL-X에서 추출한 신체 치수와 패턴 유형별로 변형된 의복 치수를 입력으로 사용하여, GarmentoPIA, GarmentoPIA†(검토기를 제외한 구조)와 baseline 시스템의 DSL 및 Python 문법 처리 성공률을 비교한다. GarmentoPIA는 프롬프트 개선 없이도 baseline보다 높은 성공률을 보였으며, 프롬프트 개선을 적용하면 두 가지 작업 유형 모두에서 우수한 성공률과 복잡한 작업에서의 현저한 성능 향상을 기록한다. 오류 분석 결과,

표 1: 시스템 간 정량평가 결과

System	Task Success Rate			Failure Types		Geometric Accuracy
	One Panel	Two Panels	All	Syntax Error	Numerical Error	mIoU↑
Baseline	30.0%	0.0%	17.1%	37.1%	45.8%	0.825
GarmentoPIA†	50.0%	33.3%	42.9%	54.2%	2.9%	0.979
GarmentoPIA	90.0%	80.0%	85.7%	4.8%	9.5%	0.975

baseline은 수치 오류 비율이 높았고, GarmentoPIA†는 수치 오류가 적으나 DSL 사용에 따른 문법 오류가 상대적으로 높다. 그러나 프롬프트 개선 방법을 통해 전체 오류율을 4.8%로 매우 감소시켰으며, mIoU 평가에서도 GarmentoPIA와 GarmentoPIA†가 baseline보다 높은 형태적 정확도를 달성하였다. 종합적으로, baseline은 빠진 패턴 구조, 잘못된 값, 문법 오류 등으로 인해 가장 낮은 성능을 보였고, GarmentoPIA†는 DSL 내 잘못된 값 할당은 적으나 빠진 패턴 구조 문제를 겪었으며, GarmentoPIA는 프롬프트 개선 방법을 통해 이러한 문제를 해결하여 가장 높은 성공률을 기록하였다 (표 1 참조).

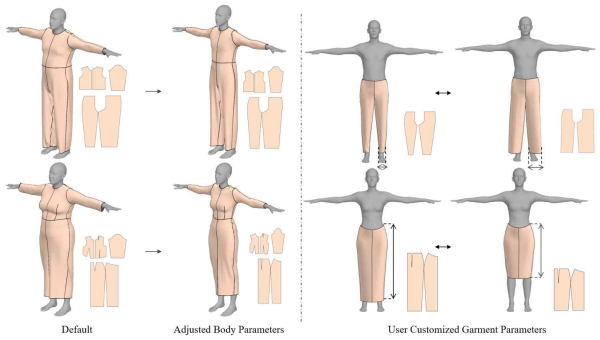


그림 3: 치수가 변경된 의복 패턴 결과

GarmentoPIA로 생성된 의복 패턴 모델은 입력된 신체 치수에 자동으로 적응하는 패턴을 생성할 수 있으며, 사용자는 의복 치수 조정을 통해 다양한 스타일을 맞춤 설정할 수 있다 (그림 3 참조). 구체적으로, 기본 의복 치수를 조정 후, SMPL-X 신체 모델의 형태 파라미터를 변경하여 목표 신체 형상을 정의하고, 해당 랜드마크로부터 신체 치수를 추출한다. 최종적으로 의복 패턴 모델이 입력된 신체 형태와 의복 사양을 모두 반영한 패턴을 자동으로 생성한다.

참고문헌

[1] KOROSTELEVA M., LEE S.-H.: Generating datasets of 3d garments with sewing patterns. *In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)* (2021).

[2] KOROSTELEVA M., SORKINE-HORNUNG O.: Garmentcode: Programming parametric sewing patterns. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–15.

[3] PEPIN H.: Modern Pattern Design: The Complete Guide to the Creation of Patterns as a Means of Designing Smart Wearing Apparel. *Funk & Wagnalls Company*, 1942.

센서 노이즈 환경에서 입력 방식에 따른 강화학습 정책의 강인성 비교

이규석[0,1] , 유리[1]
아주대학교 소프트웨어학과[1]
lgs8106@ajou.ac.kr, riyu@ajou.ac.kr

A Comparative Study on the Robustness of Reinforcement Learning Policies for Different Inputs under Sensor Noise

Gyu-Seok Yi[0,1], Ri Yu[1]
Dept. of Software and Computer Engineering, Ajou University[1]

요약

시뮬레이션에서 강화학습으로 학습된 로봇 정책을 현실 세계에 적용하는 Sim-to-Real 연구는 환경의 변화나 센서의 측정 오류와 같이, 시뮬레이션에는 없는 노이즈로 인해 성능의 저하를 겪는다. 본 연구는 이러한 현실 간극을 줄이기 위한 기초 연구로서, 어떤 시각 입력 방식이 여러 센서 노이즈에 더 강인한 정책을 생성하는지 탐구한다. 이를 위해 시뮬레이션 환경에서 로봇 팔 큐브 쌓기 과제를 대상으로, RGB 입력과 Depth 입력을 사용하는 강화학습 정책을 각각 학습시켰다. 이후, 가우시안 노이즈, 픽셀 누락 같은 실제 센서가 겪을 수 있는 다양한 노이즈를 시뮬레이션에 적용하여 얻은 결과로 각 정책의 강인성을 비교 분석한다. 실험 결과, 이상적인 환경에서는 Depth 정책이 RGB 정책보다 높은 성능을 보였지만, 노이즈 환경에서는 RGB 정책이 Depth 정책보다 더 적은 성능 하락을 보여 높은 강인성을 나타냈다. 이는 실제 운용 환경의 노이즈 유형을 고려하여 시각 입력 방식을 선택하는 것이 Sim-to-Real 성공률을 높이는 요소임을 시사한다.

1. 서론

심층 강화학습은 복잡한 로봇 제어 문제를 해결하는 강력한 도구로 부상했으며, 특히 물리 시뮬레이션 환경에서 성과를 보여주었다. 이러한 성공에 힘입어 시뮬레이션에서 학습한 정책을 실제 로봇에 이전하는 Sim-to-Real 연구[1]가 활발히 진행되고 있는데, 시뮬레이션의 완벽한 환경과 달리, 실제 로봇이 마주하는 현실 세계는 예측 불가능한 변수가 많아 시뮬레이션에서 학습된 정책이 제대로 성능을 발휘하지 못하는 현실 간극 문제가 난관으로 남아있다.

이러한 현실 간극을 유발하는 원인 중 하나는 시각 센서 데이터의 불완전성이다. 실제 환경의 카메라는 조명, 그림자, 반사, 픽셀 누락 등 정보를 정확히 측정하지 못하는 경우가 발생한다. 이처럼 각 시각 입력 방식은 고유한 노이즈 특성을 가지며, 이는 정책 안정성과 강인성에 직접적인 영향을 미친다.

* 구두발표논문
* 학부생 주저자 논문임
* 본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2022-0-01077)

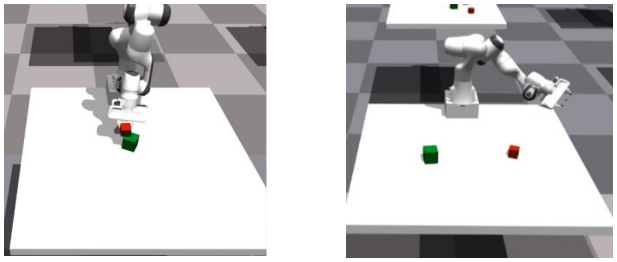
따라서 본 논문에서는 현실 간극의 문제를 완화하기 위한 연구로서, 어떤 시각 입력 방식이 현실적인 센서 노이즈에 더 강인한 정책을 생성하는지 탐구하고자 한다. 이를 위해 시뮬레이터에서 로봇 팔의 큐브 쌓기 과제를 대상으로 RGB 입력과 Depth 입력을 사용하는 강화학습 정책을 각각 학습시키고, 이후 실제 센서가 겪을 수 있는 가우시안 노이즈, 픽셀 누락 같은 노이즈를 시뮬레이션에 적용하여 각 정책의 성능 변화를 비교 분석한다.

2. 실험 설계

본 연구는 센서 노이즈 환경에서 강화학습 정책의 강인성이 시각 입력 방식에 따라 어떻게 달라지는지 비교 분석하고자 한다. 이를 위해 그림 2의 입력부에 해당하는 RGB와 Depth 입력을 각각 사용하는 두 가지 경우로 나누어 정책을 학습하고, 노이즈 유무에 따른 성능 변화를 측정하는 실험을 다음과 같이 설계했다.

2.1. 시뮬레이션 환경 및 과제

효율적인 병렬 학습을 위해 NVIDIA Isaac Gym[2] 시뮬레이터를 사용하였으며, 128개의 병렬 환경에서 학습을 동시에 진행했다. 실험 과제는 Franka Emika Panda 로봇 팔을 이용하여 테이블 위의 한 큐브를 다른 큐브 위에 쌓는 큐브 쌓기로 설정하였다 (그림 1 참조).



(a) 과제 성공 예시 (b) 과제 실패 예시

그림 1: 과제 수행 예시

2.2. 정책 학습 모델

정책 학습을 위해 Actor-Critic 모델을 사용하였으며, PPO(Proximal Policy Optimization)[3] 알고리즘을 통해 정책을 최적화하였다. 네트워크는 시각 특징을 추출하는 CNN을 액터와 크리틱이 공유하고, 이후 각자의 역할을

수행하는 두 개의 MLP로 나뉘는 구조를 가진다.

액터 네트워크: 96x96 해상도의 RGB 또는 Depth를 입력으로 받아, 합성곱 계층으로 구성된 CNN을 통해 특징을 추출한다. MLP 계층을 거쳐 로봇의 7개 관절과 2개의 그리퍼를 제어하는 9차원의 연속 행동을 출력한다.

크리틱 네트워크: 공유된 CNN의 시각 특징과 더불어, 그림2의 State에 해당하는 로봇 관절 및 큐브의 위치와 회전 정보를 입력으로 받아 현재 상태의 가치를 평가하는 단일 스칼라 값을 출력한다.

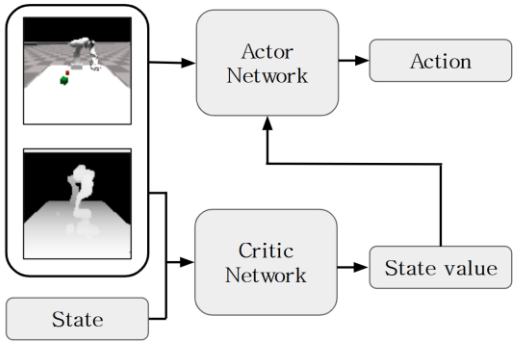


그림 2: 네트워크 구조

2.3. 센서 노이즈 모델링

현실 세계의 불완전한 센서 정보를 모사하기 위해 다음과 같은 노이즈 모델을 각 시각 입력에 적용하였다.

가우시안 노이즈: 일반적인 센서 측정 오류를 모사하기 위해 각 픽셀 값에 정규분포를 따르는 무작위 값을 추가하였다.

픽셀 누락: 센서 정보가 소실되는 현상을 모사하기 위해 특정 비율의 픽셀 값을 임의로 제거하였다.

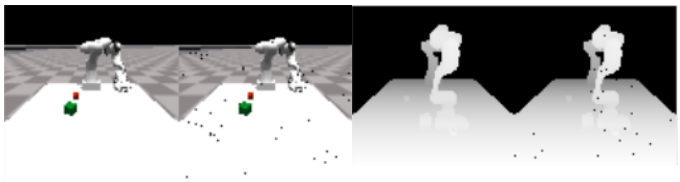


그림 3: 픽셀누락 적용 예시

3. 실험 및 결과

3.1. 강인성 평가 방법

정책의 강인성 측정을 위해, 노이즈가 없는 기준 환경에서 학습된 정책의 성능을 평가한다. 이후, 동일한 정책을 노이즈가 적용된 환경에서 테스트를 진행하여 성능의 변화를 관찰한다. 정책의 성능은 10스텝 이상 진행된 유효 에피소드에서 약 500개의 샘플을 수집하여 평균적인 보상을 계산하였고, 강인성은 기준 대비 성능 저하율로 측정하였다.

3.2. 보상 측정 방법

로봇 팔이 과제를 얼마나 성공적으로 수행하는지에 따라 여러 요소로 구성된다. 거리(그리퍼와 큐브 사이 거리에 반비례), 들어올리기(큐브가 일정 높이 이상 올라갔는지), 정렬(큐브 A가 큐브 B위에 정렬되었는지), 쌓기(큐브 A가 큐브 B위에 성공적으로 쌓였는지) 등을 고려하여 보상이 결정된다.

3.3. 실험 결과

입력방식	환경	평균 보상	성능 저하율
RGB	기준	11.59	-
	픽셀누락(0.5%)	3.69	68.2%
	픽셀누락(0.25%)	7.31	36.9%
	가우시안(0.02)	2.87	75.2%
	가우시안(0.01)	6.02	48.1%
Depth	기준	13.97	-
	픽셀누락(0.5%)	2.05	85.3%
	픽셀누락(0.25%)	3.76	73.1%
	가우시안(0.02)	1.98	85.8%
	가우시안(0.01)	2.49	82.2%

표 1: 입력 방식 및 노이즈에 따른 실험 결과

표1에서 '픽셀누락'의 괄호 안 숫자는 임의로 제거된 픽셀의 비율을 의미하며, '가우시안'의 괄호 안 숫자는 추가된 노이즈 정규분포의 표준편차를 의미한다.

4. 결론 및 한계점

Sim-to-Real 전환 과정에서 발생하는 현실 간극 문제에 대해, 본 연구는 RGB와 Depth 입력 방식의 강인성을 다양한 센서 노이즈 환경에서 비교 분석하였다. 실험 결과(표 1 참조), 노이즈가 없는 이상적인 환경에서는 기하학적 정보가 풍부한 Depth 정책의 성능(13.97)이 RGB 정책의 성능(11.59)보다 높았으나, 가우시안 노이즈나 정보 소실이 있는 픽셀 누락 같은 환경에서는 RGB 정책이 성능 저하율이 36.9%~75.2%인 반면, Depth 정책은 73.1%~85.8%에 달하는 성능 저하율을 보였다. 이는 특정 로봇 응용 분야에 센서를 선택할 때, 단순히 최고 성능에만 집중하기보다, 실제 운용 환경에서 예상되는 주요 노이즈 유형을 분석하고 이에 강인한 시각 입력 방식을 선택하는 것이 Sim-to-Real 성공률을 높이는 핵심 요소임을 시사한다. 특히, Depth 센서가 제공하는 3D 정보의 이점에도 불구하고, 실제 환경의 복잡한 노이즈 특성에 대한 민감성을 고려할 때, RGB 센서의 상대적인 노이즈 강인성이 특정 시나리오에서는 더 실용적인 대안이 될 수 있다.

하지만 본 연구는 다음과 같은 한계점을 가진다. 연산의 효율을 위해 학습에 96x96 저해상도 이미지를 사용하였다. 이로 인해 고해상도 이미지에서 나타날 수 있는 미세한 노이즈 영향이 특징 추출의 차이를 충분히 반영하지 못했을 수 있다. 또한 큐브 쌓기라는 제한된 환경에서의 단일 과제에 대해서만 실험이 이루어졌기 때문에 결과의 일반화에 한계가 있다.

참고문헌

[1] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. 2017. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. *arXiv preprint arXiv*:1703.06907, 2017

[2] V. Makoviychuk et al. 2021. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning. *arXiv preprint arXiv:*2108.10470, 2021

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv*:1707.06347, 2017

삽 기반 조작 동작과 메타 정책을 통한 사족보행 로봇의 물체 수집 전략 학습

백찬우[0], 이윤상
한양대학교 컴퓨터·소프트웨어학과
{bcw0430, yoonsanglee}@hanyang.ac.kr

Learning Object Collection Strategies for a Quadruped Robot via Shovel-Based Behaviors and Meta-Policies

Chanwoo Baek[0], Yoonsang Lee
Dept. of Computer Software Engineering, Hanyang University

요약

본 연구는 사족보행 로봇에 삽(shovel)을 장착하여 바닥에 있는 물체를 짐칸에 수집하고, 이후 몸을 기울여 물체를 지정된 수집 통에 옮기는 연속 동작을 학습하는 것을 목표로 한다. 이전 연구에서는 삽을 통한 수집까지만 가능했으나[1], 본 연구에서는 로봇의 자세 조절을 통해 수집한 물체를 외부 공간으로 정확히 투하하는 동작까지 포함한 트랜스포팅 루틴(transporting routine)을 학습한다. 이를 위해 접근(π_approach), 수집(π_st), 투하(π_dump)의 세 정책을 메타 정책(π_meta, meta policy)를 통해 유기적으로 전환하며, 시뮬레이션 기반의 강화학습 환경에서 이를 실현한다.

1. 서론

기존의 로봇 기반 물체 수집 및 이송 작업은 주로 로봇 팔과 같은 정밀 조작 장비를 필요로 하며[2], 에너지 효율성과 기계적 복잡도 측면에서 한계가 있다. 본 연구에서는 이러한 한계를 극복하고자 사족보행 로봇에 삽 구조를 부착하고, 사지의 조작만으로 바닥에 있는 물체를 수집하고 외부 통에 이송하는 연속적 조작 루틴(Sequential Manipulation Routine)을 강화학습 기반으로 학습시킨다. 특히 이전 연구에서는 삽을 이용해 물체를 짐칸에 싣는 작업까지만 가능했으나[1], 이번 연구에서는 그 후속 동작으로 몸을 기울여 물체를 외부 수집 통에 붓고, 자세를 복원한 뒤 다시 다음 물체를 수집하러 이동하는 일련의 절차를 수행한다. 이를 위해 동작별로 구분된 세 가지 정책(접근, 수집, 투하)을 구성하고, 상황에 따라 적절한 정책의 동작을 수행하도록 하는 메타 정책 구조를 설계하였다.

* 구두(포스터) 발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회 (논문지)에 제출 준비중임.
* 본 연구는 xxx 지원으로 수행되었음.

2. 제안하는 방법

2.1. 정책 구조

본 연구에서는 사족보행 로봇의 연속적 물체 운반 루틴(Sequential Transporting Routine)을 구현하기 위해, 세 가지 개별 정책을 구성하고 이들을 메타 정책 구조로 통합하는 방식을 설계하였다.

- π_approach: 로봇이 지정된 목표 지점에 접근하도록 학습된 정책으로, 이동 중 로봇의 정면이 진행 방향과 정렬되도록 요(yaw)를 조절하며 전진한다. 따라서 π_approach는 로봇이 이동 경로를 따라가며 자신의 방향(orientation)을 조정할 수 있도록 한다.
- π_st(scoop-toss): 물체를 쳐서 삽 안쪽으로 굴려 담은 뒤, 다리 관절의 빠른 움직임을 통해 물체를 짐칸에 던지는 정책이다. 물체 던지기 동작은 고정된 삽에 대해 로봇 관절의 움직임을 조정하여 수행되며, 이를 통해 물체가 짐칸 방향으로 정확히 투사되도록 유도한다.
- π_dump: 로봇의 상체를 들어올리고 뒷부분을 낮추는 동작을 통해 몸을 뒤로 기울여, 짐칸의 물체를 수집 통에 붓는 정책이다. 물체 낙하를 인식하면 로봇은 기본 자세로 복원하며 루틴을 반복한다.

이러한 세 정책은 향후 메타 정책 구조 하에서 상황에 따라 전환되며, 전체 운반 루틴을 수행하게 된다.

2.2. 단계별 학습 과정

본 연구에서는 전체 운반 루틴 중에서도 π_dump 정책

의 안정적인 학습을 우선적으로 구현하였으며, 이를 위해 2단계 학습 방식을 도입하였다. 이와 같은 학습 분할은 초기 단계에서의 동작 안정성 확보와 후속 동작의 효율적인 수렴을 유도하기 위한 것이다. 본 절에서는 현재까지 중점적으로 다룬 π_dump의 학습 과정과 이에 적용된 리워드 구조를 설명한다.

1단계에서는 로봇이 짐칸에 담긴 물체를 담은 채 후방으로 이동하는 동작을 먼저 학습한다. 이러한 후방 이동 동작의 학습 과정을 통해 피치(pitch)를 올리는 방향의 움직임을 유도하고, 기울어짐에 대한 기본 동작 감각을 익히도록 설계하였다.

2단계에서는 상체를 들어올려 피치를 증가시키는 동작을 학습하며, 실제 물체를 투하하기 위한 자세를 형성하는데 집중한다.

2.3. 보상(reward) 구성

본 절에서 설명하는 보상 구조는 현재까지 구현 및 학습이 진행된 π_dump 정책에 적용된 부상 구조를 중심으로 한다.

1단계 부상은 로봇이 후방으로 이동하는 동작을 먼저 학습하도록 유도하기 위해, 이동 거리 증가에 따른 선형 보상으로 구성되었다. 이 단계의 목적은 피치 자세 자체를 학습하는 것이 아니라, 2단계에서 피치 자세 탐색이 더 효과적으로 이루어질 수 있도록, 초기 자세에서의 물리적 진입 경로를 확보하는 것에 있다. 실제로 로봇이 평평한 자세에서 곧바로 피치를 크게 올리려 하면 실패 확률이 높고 탐색이 제한되지만, 후진 동작을 통해 로봇이 뒤로 이동하는 관성이 생긴 상태에서는 기울임 동작이 더 자연스럽고 안정적으로 시작될 수 있다.

2단계 보상은 1단계에서 학습된 후진 동작이 정책의 초기 행동에 영향을 주어, 피치 기울임에 유리한 탐색 경로를 자연스럽게 유도한다는 점을 활용하여 설계되었다. 이 단계에서는 로봇이 상체를 들어올려 피치 값을 크게 증가시키고, 일정 시간 동안 그 기울기를 유지하는 동작을 학습한다. 피치는 0°(수평)에서 −90°(수직)에 해당하는 범위를 0~1 사이로 정규화하고, 이를 45개의 구간으로 균등하게 분할하였다. 각 구간에 도달할 때마다 보너스(bonus) 보상이 점진적으로 증가하도록 설계되어, 더 깊은 기울기에 도달할수록 더 큰 보상이 주어진다. 보너스 보상은 구간의 인덱스(index)를 기반으로 선형으로 증가하며, 이를 통해 높은 피치 각도 도달을 강하게 유도하였다. 또한 목표 피치 구간에 도달한 이후에도 불안정한 자세로 인해 다시 범위 밖으로 이탈하는 행동을 억제하기 위해, 직전 스텝과 비교했을때 탈락하는 구간이 있을 경우 해당 구간에서 제공되던 보너스의 1.1배에 해당하는 값을 보너스 보상에서 빼는 패널티(penalty)를 적용하였다.

3. 실험

본 연구에서는 π_dump 정책의 학습 성능과 보상 구조의 효과를 검증하기 위해 Isaac Gym(아이작 짐) 기반 시뮬레이션 환경에서 실험을 수행하였다.

기존 연구[1]에서는 π_approach와 π_st를 연계한 메타 정책을 학습하여 반복적인 수집 루틴을 구성하였으나, 본 연구는 후속 정책인 π_dump 단일 동작의 안정적인 수행에 초점을 맞추었다. 학습된 π_dump 정책은 로봇이 정지 자세에서 상체를 기울여 물체를 외부 수집 통에 투하하는 동작을 성공적으로 수행하며, 이 과정을 시각화한 결과를 그림 1에 정리하였다.

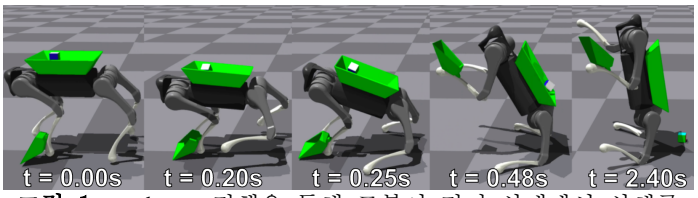


그림 1: π_dump 정책을 통해 로봇이 정지 상태에서 상체를 기울여 물체를 붓는 일련의 동작
(t는 기준 시점으로부터의 경과 시간(초))

4. 결론 및 향후 연구

본 연구에서는 사족보행 로봇이 물체를 짐칸에 담은 뒤 몸을 기울여 외부 수집 통에 투하하는 동작을 학습하는 π_dump 정책을 설계하고, 이를 Isaac Gym(아이작 짐) 기반 시뮬레이션 환경에서 구현하였다. π_dump는 기존 연구에서 제안된 π_approach 및 π_st 정책과 연계되어 전체 운반 루틴을 구성하는 메타 정책의 후속 동작으로 작동한다. 메타 정책을 통해 세 가지 하위 정책 간 전환이 자동으로 이루어지며, 이 중 π_dump는 루틴의 종단 동작으로서 작업의 완성도를 결정짓는 중요한 역할을 수행한다.

향후 연구에서는 우리의 시뮬레이션 기반 정책 전환을 실물 환경에서의 정책 전환으로 확장할 계획이다. 이를 위해 RGB 카메라와 비전 트랜스포머(ViT, Vision Transformer) 인코더를 활용하여 물체의 정보를 추정할 수 있도록 설계하기를 구상중이다. 이러한 후속 연구를 통해, 사족보행 기반 시스템의 활용도를 한층 더 향상시킬 수 있을 것으로 기대된다.

참고문헌

[1] M. Kang, C. Baek, and Y. Lee, Scoop-and-Toss: Dynamic Object Collection for Quadrupedal Systems, arXiv preprint arXiv:2406.12345, 2024.
[2] F. Zhang, L. Xiao, Y. Wang, and R. Xiong, Deep Whole-Body Control: Learning a Unified Policy for Manipulation and Locomotion, in Proc. Conf. Robot Learning (CoRL), 2022.

저품질 3D 모델의 텍스처 및 기하 품질 향상*

류누리[1], 원지윤[2], 손주은[2], 공민수[2], 이주행[3], 조성현[12]
포항공과대학교 {인공지능대학원[1], 컴퓨터공학과[2]}, 페블러스[3]
{ryunuri, w1jyun, jeson, gongms}@postech.ac.kr, joohaeng@pebblous.ai, s.cho@postech.ac.kr

Elevating 3D Models: High-Quality Texture and Geometry Refinement from a Low-Quality Model

Nuri Ryu[1], Jiyun Won[2], Jooeun Son[2], Minsu Gong[2], Joo-Haeng Lee[3], Sunghyun Cho[123]
POSTECH {GSAI[1], CSE[2]}, Pebblous[3]

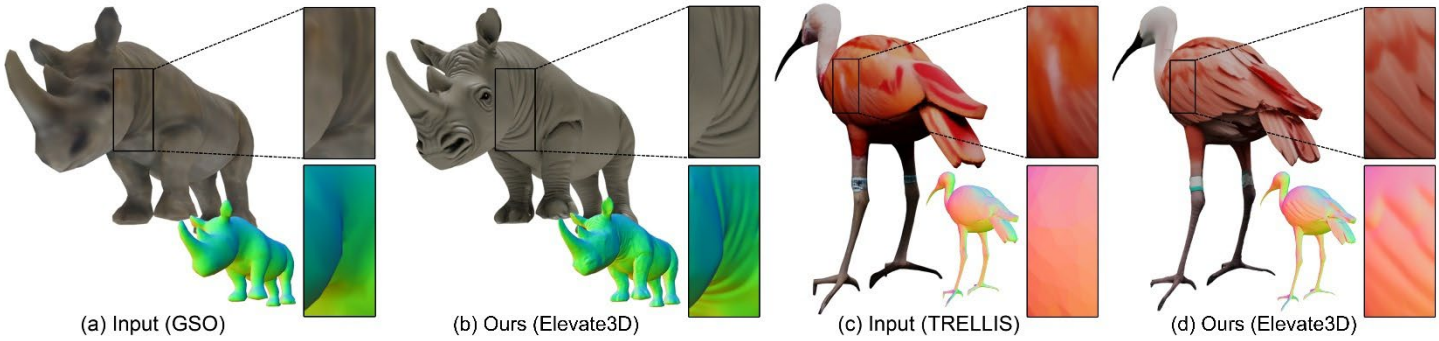


Figure 1: 3D refinement examples from (a) a degraded real-world scan and (c) a state-of-the-art image-to-3D generative model. Elevate3D effectively refines both texture and geometry while preserving their alignment, as shown in (b) and (d).

Abstract

Elevate3D enhances low-quality 3D models into high-quality assets. It employs an iterative process where HFS-SDEdit first performs identity-preserving generative texture refinement using high-frequency guidance. This refined texture subsequently guides geometry improvement, ensuring high-quality 3D results with well aligned texture and geometry.

1. Introduction

High-quality 3D models are in unprecedented demand but remain scarce due to high acquisition costs. This motivates refining easily accessible low-quality models to achieve higher quality. Recent 3D model refinement methods leverage generative diffusion priors, often employing techniques like SDEdit [1] for texture refinement. However, these approaches often suffer from key limitations. First, SDEdit's core mechanism, which involves adding noise to an input and then denoising it, inherently creates a quality-fidelity trade-off controlled by the noise level. Second, the predominant reliance on image-based generative prior often leaves geometry under-constrained or misaligned with textures.

This paper proposes Elevate3D, a novel 3D model refinement approach that produces a high-quality 3D model with well-aligned texture and geometry through iterative, view-by-view refinement. Elevate3D first enhances texture using High-Frequency-Swapping SDEdit (HFS-SDEdit), which resolves SDEdit's trade-off by using high-frequency guidance to preserve input identity while allowing for high quality refinement. Subsequently, geometry refinement is guided by the normal maps predicted from the refined texture, ensuring alignment of texture and geometry.

2. Method

Elevate3D refines 3D models through iterative, view-by-view texture and geometry refinement stages. Texture refinement leverages HFS-SDEdit, which builds upon SDEdit [1] to overcome its inherent quality-fidelity trade-off. HFS-SDEdit employs high-frequency guidance from the input. This allows the

---

* 구두 발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 SIGGRAPH 2025에 발표 예정.
* 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-01906, 인공지능대학원 지원(포항공과대학교))과 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2024-00457882, 인공지능 혁신 허브 연구 개발)과 페블러스의 지원을 받아 수행된연구임

Figure 2: HFS-SDEdit addresses the quality-fidelity trade-off in SDEdit. By adding a substantial amount of noise ϵ to the low-quality reference image z_r in (c) and initiating the denoising process from the noisy latent z_{t_h}, SDEdit removes domain information, enabling the diffusion model to generate a high-quality image as depicted in (b). However, this approach compromises fidelity to the reference image. Conversely, adding a small amount of noise and starting the denoising process from the noisy latent z_{t_l} preserves the low-quality domain information, resulting in only minor refinements, as seen in (d). In contrast, HFS-SDEdit incorporates high-frequency feature injection-based guidance allowing for high-fidelity generation even when starting the denoising process from z_{t_h}. This approach achieves both high quality and fidelity in the refinement.

diffusion model to freely generate low-frequency features for high output quality, while simultaneously preserving the original identity without transferring its artifacts. We provide an overview of HFS-SDEdit in Figure 2. HFS-SDEdit initializes a noisy latent z_{t_h} using large noise from the reference z_r. Then, at subsequent timesteps t to t_{stop}, it replaces the high-frequency component of the denoised latent $\hat{z}_t$ with that from a noised reference $\tilde{z}_t$ to form a calibrated latent z'_t. Within Elevate3D, for each view, this process is applied to unrefined regions of the rendered image I_i identified by a mask m_i. The resulting refined latent z'_t is blended with $\hat{z}_t$ to preserve already refined areas, yielding an improved texture I'_i. The geometry refinement stage then utilizes this refined texture I'_i. The refined texture I'_i introduces new geometric details generated during its refinement. A normal map n_i is inferred from I'_i to capture these details. Our regularized normal integration scheme then integrates these details into the existing mesh M_i. It estimates a refined surface S_i by minimizing an energy functional $E(z)$ designed to ensure the new surface both reflects the normals n_i from the refined texture and remains consistent with the geometry of M_i. The refined geometry S_i is integrated into M_i using Poisson surface reconstruction [2], creating $\widetilde{M}_i$. Finally, the texture I'_i is projected onto this updated mesh and we proceed with subsequent iterations.

3. Experiments

We evaluated the 3D model refinement quality of Elevate3D on a degraded GSO dataset [3] against SoTA 3D refinement methods. Qualitatively, Elevate3D produced detailed textures and geometries, substantially surpassing previous methods and exhibited high texture-geometry consistency. We demonstrate the results in Figure 1. Quantitatively, Elevate3D consistently outperformed competitors on non-reference metrics for rendered images.

For evaluating HFS-SDEdit, experiments showed that swapping latent low-frequencies from a low-quality reference degraded output quality, whereas swapping only its high-frequencies allowed the diffusion model to maintain high output quality. This confirmed that low-frequency components primarily carry image-quality domain information, justifying HFS-SDEdit's strategy of using high-frequency guidance from the input for identity preservation while allowing the model to synthesize high-quality low-frequency features. In image refinement comparisons on the LSDIR dataset against other SDEdit based methods, HFS-SDEdit achieved the best performance in non-reference metrics and LPIPS, producing high-quality outputs with convincing fidelity. Finally, ablation studies demonstrated the necessity of both texture and geometry refinement stages and our regularized normal integration for high-quality results.

4. Future Work

While Elevate3D produces high-quality textured meshes, its processing time currently scales with the number of refinement views. We plan to optimize this speed in future work while preserving quality.

References

[1] Meng et al., SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, ICLR, 2022
[2] Kazhdan et al., Poisson surface reconstruction, SGP, 2006
[3] Downs et al., Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items, ICRA, 2022

단안 비디오를 이용한 야구 피칭 모션 재건*

김지원[0,1], 유리[1,2]
아주대학교 인공지능학과[1], 아주대학교 소프트웨어학과[2]
jw731@ajou.ac.kr, riyu@ajou.ac.kr

Baseball Pitching Motion Reconstruction from Single-View Videos

Ji-won Kim[0,1], Ri Yu[1,2]
Dept. of Artificial Intelligence, Ajou University[1]
Dept. of Software and Computer Engineering[2]

Abstract

We propose a two-stage learning framework for reconstructing 3D baseball pitching motions from video, addressing challenges such as motion blur in high-speed throws. Our approach combines motion imitation and refinement using physics-based simulation and deep reinforcement learning, guided by domain-specific rewards. Despite imperfect pose estimates, our method generates plausible, physically consistent motions that preserve pitching style and accurately replicate ball trajectory and speed.

1. Introduction

Reconstructing 3D human motion from videos is challenging due to motion blur and self-occlusion, especially in fast sports like baseball pitching. While prior methods [1,2] focused on correcting pose errors at the image level, we propose a physics-based approach that reconstructs physically consistent motions directly through simulation. Our two-stage framework, combining physics-based simulation and deep reinforcement learning [3], refines motion imitation using task-specific rewards, ensuring realistic pitching even with degraded input data. We demonstrate its effectiveness through ablation studies and comparisons, showcasing high-quality 3D motion recovery from single-view videos with blur and occlusion.

* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회(논문지)에 제출 중.
* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구결과로 수행되었음 (IITP-2025-RS-2023-00255968)

2. Method

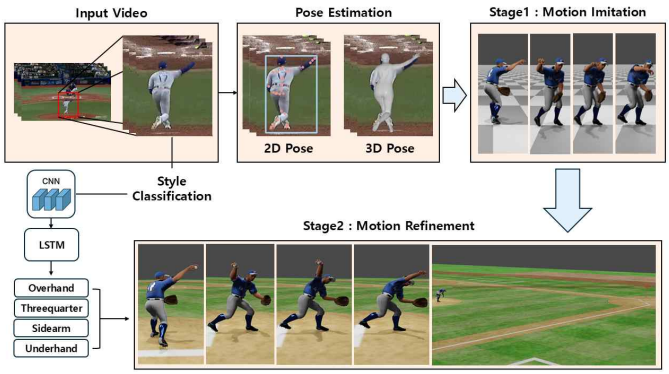


Figure 1: System overview

2.1. Motion Reconstruction

Baseball pitching is a highly dynamic motion often degraded in video-based reconstruction due to motion blur and self-occlusion. Prior methods that rely solely on pose imitation struggle under such conditions. To address this, we incorporate domain-specific knowledge—such as target accuracy, ball speed, and pitching style—as reward signals in reinforcement learning to guide the reconstruction process. Using deep reinforcement learning and physics-based simulation, we design task-specific rewards that encourage realistic ball delivery. To resolve conflicts between pose imitation and task objectives, we adopt a two-stage approach: the first stage focuses on imitating reference poses, while the second refines the motion based on task constraints. This sequential training enables accurate and physically plausible reconstructions, even from noisy pose data.

2.2. Two-Stage Reinforcement Learning for Motion Reconstruction

To reconstruct realistic baseball pitching motions

from noisy pose estimations, we adopt a two-stage reinforcement learning framework. In the first stage, the goal is to imitate reference poses extracted from video frames, despite their imperfections due to motion blur and self-occlusion. This stage focuses on tracking joint angles, joint velocities, end-effector positions, and foot contact states, while also applying torque regularization to ensure physical plausibility. To account for inaccuracies in the pitching arm caused by motion blur, its tracking weight is reduced.

In the second stage, we refine the motion using domain-specific knowledge about baseball pitching. This includes encouraging the character to throw the ball toward a predefined target location, align the ball's release direction with the strike zone, and match the throwing speed extracted from the video. Additionally, we introduce a pitching style reward based on the arm's release angle, which is classified into four types: overhand, three-quarter, sidearm, and underhand. By aligning the simulated arm angle with the identified pitching style, we ensure that the reconstructed motion maintains stylistic fidelity to the original video. This two-stage process enables the system to balance imitation and task performance, resulting in physically consistent and visually realistic pitching motions.

3. Result

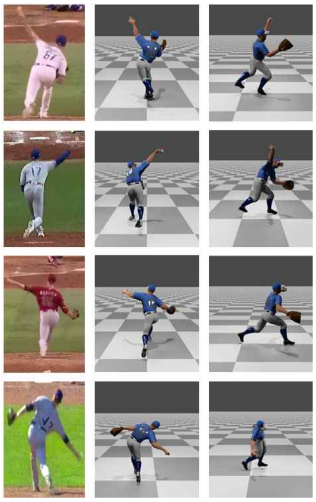


Figure 2: Results of various pitching styles

To evaluate the effectiveness of our proposed framework, we conducted a series of experiments. First, we demonstrate that the system can reconstruct realistic pitching motions from video clips featuring various throwing styles, including overhand, three-quarter, sidearm, and underhand. The reconstructed motions successfully reflect the distinctive posture and dynamics at the moment of ball release, closely matching the original video appearances.

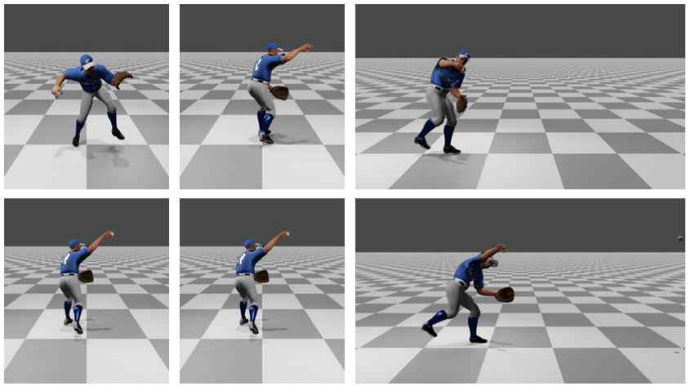


Figure 3: Stage-wise performance comparison results

We also assessed the impact of our two-stage training strategy. In the first stage, the model focuses on imitating the overall motion, while in the second stage, it refines the motion using baseball-specific cues such as ball speed, target direction, and pitching style. The comparison reveals that motions refined in the second stage display more dynamic and physically plausible movements—especially in stride length and arm action—than those generated after only the first stage.

4. Conclusion

We propose a robust framework for reconstructing complex sports motions, such as baseball pitching, from monocular video, even under challenging conditions like motion blur. By combining deep reinforcement learning with physics-based simulation in a two-stage process—imitation and refinement—our method produces high-quality, realistic motions. This makes it suitable for applications in sports analysis, motion tracking, and virtual training, particularly when input videos are degraded.

Reference

[1] Bright J, Chen Y, Zelek J. Mitigating Motion Blur for Robust 3D Baseball Player Pose Modeling for Pitch Analysis. arXiv preprint arXiv:2309.01010. 2023.

[2] Zhao Y, Rozumnyi D, Song J, Hilliges O, Pollefeys M, Oswald MR. Human from Blur: Human Pose Tracking from Blurry Images. arXiv preprint arXiv:2303.17209. 2023.

[3] Yu R, Park H, Lee J. Human Dynamics from Monocular Video with Dynamic Camera Movements. ACM Trans. Graph.. 2021;40(6).

포스터 발표

확장현실 기반 의료기기 실습 교육 도구 개발

김영서[0], 전홍익, 최승관*, 박상훈*
서강대학교 메타버스전문대학원
(dudtj190, rhavkd33, csk0123, mshpark)@sogang.ac.kr

Development of Extended Reality-based Training Tools for Medical Equipment

Young-Seo Kim[0], Hong-Ik Jeon, Seung-Gwan Choi*, Sang-Hun Park*
Sogang University Graduate School of Metaverse

요약

본 연구는 전통적 의료 교육의 한계를 보완하기 위해 개발된 확장현실 기반 X-ray 실습 교육 도구의 효과를 확인한다. 확장현실 환경에서 핸드트래킹 기반 상호작용을 구현하고, 학습 흥미도, 숙련도 향상, 지속 사용성을 평가했다. 이를 통해 확장현실 기반 의료기기 실습 교육은 학습 흥미도와 숙련도 향상에 긍정적 효과가 있음을 확인했다. 본 연구는 확장현실 기술이 의료 교육에서 실질적이고 효율적인 대안이 될 수 있음을 시사하며, 향후 다양한 의료기기 실습 시나리오에서의 적용 가능성을 제시한다.

1. 서론

의료기기 기술의 고도화로 조작 방법이 복잡해짐에 따라 실습 기반 의료기기 전문 교육의 필요성이 증가한다. 그러나 공간적 제약, 방사선 노출 위험 등의 한계로 의료기기 실습 교육이 제한되고 있다 [1, 2, 3]. 때문에, 최근 이런 한계를 극복하기 위한 대안으로 확장현실 기술이 주목받고 있으며 [4], 특히 핸드트래킹(hand tracking) 기반 상호작용은 현실과 유사한 경험을 제공한다 [5]. 본 연구는 확장현실 기반 의료기기 실습 교육 도구를 개발하고, 학습 흥미도, 숙련도 향상, 지속 사용성 측면에서의 교육 효과를 검증하고자 한다.

2. 실습 교육 도구 구현

본 연구는 방사선 노출 위험이 존재하며 무거워 전통적 의료 교육에서 쉽게 이용할 수 없는 의료기기인 X-ray를 실습 교육 도구 모델로 선택했으며, 실습할 상용 장비로 GE Healthcare Definium XR/f를 선택해 확장현실 기반 의료기기 실습 교육 도구를 개발했다.

(a) Rotate tray (b) Move table (c) Drag telescopic arm

Figure 1: Hand tracking based interaction

확장현실 기반 의료기기 실습 교육 도구는 Figure 1처럼 사용자가 확장현실에서 본인의 손으로 가상의 X-ray 장비를 잡고 움직이는 등의 상호작용을 할 수 있다. 이를 통해 실습자는 실제 장비를 조작하듯이 가상의 모델을 조작하며 실습할 수 있다.

(a) Virtual tablet interface (b) Video manual

Figure 2: Virtual tablet interface based education

또한 실습 교육 도구에서 제공하는 가상 태블릿 인터페이스(virtual tablet interface)는 Figure 2와 같이 의료기기 이용 방법을 설명하는 동영상 기반 매뉴얼, 각 환자 자세 별 실습을 진행할 수 있는 실습 시작 버튼으로 구성되어 있다. 사용자는 왼손의 엄지와 검지를 빠르게 두 번 부딪쳐 가상 태블릿 인터페이스를 호출한 뒤, 동영상 매뉴얼을 시청하고, 자세 별 실습 시작 버튼을 클릭해 각 환자 자세 별 의료기기 이용 실습을 진행할 수 있다.

3. 사용성 평가

3.1 사용성 평가 지표

교육 효과를 검증하는 평가 지표로 5점 리커트 척도(Likert scale) 기반 설문조사와 심층 인터뷰를 선정했다. 설문조사는 (1) 학습 흥미도, (2) 숙련도 향상, (3) 지속 사용성에 대해 평가하는 문항으로 구성되어 있다. 심층 인터뷰는 확장현실 기반 의료기기 실습 교육 도구 이용 경험과 긍정적인 부분, 아쉬운 부분을 심층적으로 질문하는 문항으로 구성했다.

3.2 실험 참여자

실험 참여자로 강릉원주대학교 간호학과 학생 8명, 연세의료원 디지털 헬스 전략실 메타버스 프로젝트에 참여한 경력을 보유한 세브란스병원 외래간호팀 간호사 1명, 5회 이상 확장현실 기반 실습 경력을 보유한 서울성모병원 기능검사팀 임상병리사 1명을 섭외했다.

3.3 사용성 평가 결과

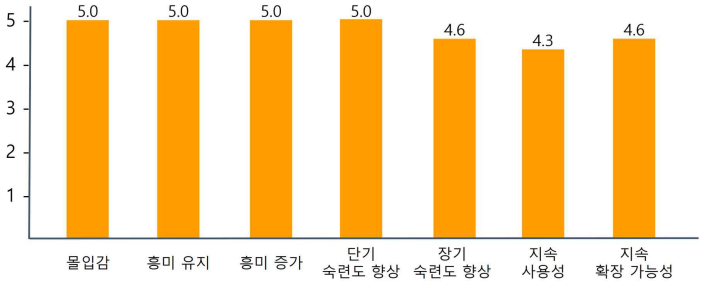


Figure 3: Result of Likert scale based research

간호학과 학생을 대상으로 한 설문조사 결과, 여러 측면에서 긍정적인 평가를 받았다. 특히, 몰입감, 흥미 유지, 흥미 증가, 그리고 단기 숙련도 향상은 모두 5.0점을 기록하며 높은 만족도를 나타냈다. 이는 방사선 노출 위험 없이 안전하게 학습할 수 있으며, 손을 이용한 상호작용이 현실과 유사한 환경을 제공한다는 점에서 큰 이점으로 작용한 것이다. 장기 숙련도 향상과 지속 확장 가능성 또한 각각 4.6점으로 높은 평가를 받았다. 반면, 지속 사용성 항목은 4.3점으로 상대적으로 낮은 점수를 기록했는데, 이는 현 실습의 다양성이 부족하다는 점이 아쉬움으로 작용한 결과이다(Figure 3 참고).

세브란스병원 외래간호팀 간호사와 서울성모병원 기능검사팀 임상병리사는 공간 제약 해소, 자연스러운 상호작용을 긍정적으로 평가했으며, 실제 장비 기반 실습 대비 숙련도 향상과 확장 가능성에서 강점을 보인다고 분석했다. 이들은 해당 도구가 다양한 의료기기 실습과 시뮬레이션 환경에서 폭넓은 적용이 가능할 것이라는 의견을 제시했다.

4. 결론

확장현실 기반 의료기기 실습 교육 도구는 학습자에게 학습 흥미도와 숙련도 향상에 긍정적인 효과를 미친다. 하지만 소수의 교육 도구를 개발해 지속 사용성 부분에서 큰 효과를 발휘하지 못한다. 의료기기 실습 교육 전문가를 대상으로 한 심층 인터뷰는 해당 도구가 실습의 숙련도 향상에 큰 강점을 가짐을 확인한다. 특히 방사선 노출 위험이 없고, 공간적 제약이 없는 환경에서 실제와 비슷한 상호작용할 수 있다는 점에서 실습 교육 도구로서의 가치를 평가받았다. 동시에 의료기기 실습 도구 및 실전 시뮬레이션으로의 확장 가능성이 높음을 보여준다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이고(RS-2023-00251681), 정보통신기획평가원의 대학ICT연구센터사업(RS-2023-00259099)과 메타버스융합대학원(RS-2022-00156318)의 지원으로 수행되었음.

References

[1] 이윤지, 의료기기산업 교육 현황 분석을 통한 전문인력 양성방안. 국내석사학위논문 연세대학교 대학원, 2021. 서울.

[2] 한창화, 전영황, 한재복, 공창기, 송종남. (2023). 교육용 의료방사선 시뮬레이터 시스템 개발 및 연구 모델 제안. 한국방사선학회논문지, 17(3), 459-464.

[3] 서정민. (2021). 방사선발생장치 교육을 위한 시뮬레이터의 개발과 유용성 평가. 방사선기술과학, 44(6), 591-597. 10.17946/JRST.2021.44.6.591

[4] 김정상. (2022). 확장현실 적용 가능한 NCS 능력단위분석 -의료기기 산업을 중심으로-.한국방송통신대학교 대학원.

[5] 성현경, 신나민. (2024). 가상현실 (VR) 및 증강-현실 (AR) 기반 의료 시뮬레이션 교육에 관한 연구 동향. 대한한방소아과학회지, 38(1), 78-87.

실시간 렌더링 환경에서 Vulkan을 활용한 배치 렌더링 기반 드로우 콜 최소화

오정식[0], 박정용, 이현규†
인천대학교 컴퓨터공학부
{ojs0331[0], jylove0515, hyeonkyulee†}@inu.ac.kr

Draw Call Minimization via Batch Rendering In Real-time Rendering Using Vulkan

Jungsik Oh[0], Jeongyong Park, Hyeonkyu Lee†
Department of Computer Science and Engineering, Incheon National University

요약

본 논문은 실시간 그래픽스 파이프라인에서 빈번하게 발생하는 드로우 콜(Draw Call)로 인해 발생하는 CPU-GPU 통신 병목을 최소화하기 위한 효율적인 배치 렌더링(Batch Rendering) 기법을 제안한다. 제안된 기법은 Vulkan의 Indirect 드로우 콜 기능을 활용하여 다수의 객체를 하나의 배치 단위로 결합하여 단 한 번의 드로우 콜 호출만을 수행함으로써 반복적인 드로우 콜 호출에 따른 명령어 송신 횟수를 획기적으로 절감하였다. 또한, FPS는 평균적으로 약 22.7% 향상되었다.

1. 서론

실시간 3D 그래픽스 엔진에서는 매 프레임마다 수천에서 수십만 건에 이르는 드로우 콜이 실행된다. 각 드로우 콜은 CPU가 GPU에게 렌더링과 관련된 명령 정보를 전달하는 단위로서, 드로우 콜이 많아질수록 CPU-GPU 간 통신 오버헤드가 기하급수적으로 증가한다[1]. 이러한 병목 현상은 고해상도 및 고주사율을 지향하는 현대 게임 및 시각화에서 전체적인 렌더링 성능을 결정짓는 주요 요인으로 작용한다.

본 연구에서는 Vulkan[2]의 Indirect 드로우 콜[3]을 활용하여 사전에 설정된 하이퍼 파라미터를 기준으로 다수의 객체들을 결합한 배치 단위로 만들고 Indirect, 정점(Vertex), 인덱스(Index) 버퍼들을 생성한다. 이를 기반으로 N개의 드로우 콜을 단 하나의 Indirect 드로우 콜 호출로 통합 실행함으로써, 반복적인 명령 전달 횟수를 효과적으로 최소화한다. 이와 같은 최적화는 다수의 객체가 존재하는 복잡한 장면에서도 CPU-GPU 간 병목을 완화하고, 실시간 렌더링의 안정성과 성능을 확보하는 데 유효하다. 실제 실험에서는 제안 기법이 기존 렌더링 파이프라인 대비 드로우 콜 호출 횟수를 최대 97% 이상 절감하였으며, 평균 약 22.7%의 FPS(Frame Per Second) 향상을 달성하였으며, 이는 본 기법이 실효성이 있음을 보여준다.

2. 방법

2.1. 렌더링 파이프라인

그림 1: 배치 렌더링 파이프라인

본 연구에서 구성한 렌더링 파이프라인의 전체 절차를 그림 1에 제시하였다. 초기 배치 구축 단계는 오프라인에서 CPU에 의해 수행된다. 이 과정에서는 렌더링 대상 객체들을 사전 정의된 기준에 따라 배치 단위로 그룹화하고, Indirect 드로우 콜 호출을 위한 버퍼들을 생성한다.

2.2. 배치 및 Indirect 커맨드 버퍼 생성

배치 데이터 갱신 단계에서 CPU가 Vulkan의 Indirect 드로우 콜 호출에 필요한 다섯 개의 32비트 정수를 하나의 구조체 형태를 통해 작성한 후, 이를 임시 버퍼에 누적한다. 임시 버퍼의 크기가 사전에 설정된 하이퍼 파라미터 용량에 도달하면, 누적된 버퍼를 GPU에 전달하여 하나의 Indirect 커맨드 버퍼를 확정하고, 대응되는 정점 및 인덱스 데이터를 GPU 메모리에 할당한다.

기존 파이프라인에서는 메시(Mesh) M개마다 상응하는 정점 및 인덱스 버퍼를 개별 할당되었으며, 반면 제안하는 배치 기반 접근은 이 과정을 배치 수 $N(N \ll M)$만큼으로 집약하여, 메모리 할당 호출 횟수를 $O(M)$에서 $O(N)$으로 축소시킨다. 이를 통해 GPU 드라이버 레벨에서의 메모리 관리 오버헤드가 효과적으로 감소되며, 전

* 포스터 발표논문
* 학부생 주저자 논문
† 교신저자

체 렌더링 효율성이 향상된다.

알고리즘 1: 배치 및 버퍼 생성 과정

```
1. procedure BUILD_BATCH(meshList, Capacity)
2.     batch ← new MiniBatch()
3.     for each mesh in meshList do
4.       vBytes ← mesh.vertexCount × sizeof(Vertex)
5.       iBytes ← mesh.indexCount × sizeof(uint32)
6.       if batch.size + vBytes + iBytes > Capacity then
           # 3개의 버퍼 생성 및 새로운 배치 생성
7.         createBuffers(batch)
8.         batch ← new MiniBatch()
9.       end if
          # Draw Commands 기록
10.      appendDrawCmd(batch, mesh)
         # 정점·인덱스 정보 누적
11.      appendData(batch, mesh.vertices, mesh.indices)
12.      batch.size ← batch.size + vBytes + iBytes
13.    end for
     # 배치의 잔여 데이터에 대한 버퍼 생성
14.  if batch.size > 0 then
15.    createBuffers(batch)
16.  end if
17. end procedure
```

배치에 포함된 모든 메시의 정점 및 인덱스 데이터는 단일 연속 메모리 블록으로 결합되어 VRAM에 저장된다. Indirect 커맨드 버퍼에 기록된 각 요소 i는 동일 배치 내 정점 버퍼 및 인덱스 버퍼에서 특정한 오프셋 주소를 참조함으로써, 개별 메시의 렌더링에 필요한 메모리 영역을 정확히 지정할 수 있다.

2.3. Indirect 드로우 콜 호출

배치 기반 Indirect 렌더링의 핵심은 드로우 콜의 총 횟수를 최소화하는 데 있다. 제안한 알고리즘 1에 기술된 절차를 통해 생성된 N개의 배치로 구성된다. 배치마다 한 번의 Indirect 드로우 콜을 호출만을 수행한다. 기존 방식이 메시 수(M)만큼 발생시키던 드로우 콜을 제안 기법은 배치 수($N \ll M$)로 압축한다. 드로우 콜 호출 량은 $O(M)$에서 $O(N)$으로 감소된다.

3. 실험 및 결과

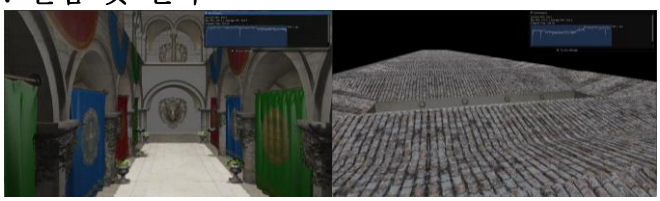


그림2: (좌)시점 변경 전 화면, (우)시점 변경 후 화면

삼각형 수($\times 10^3$)	기본(Basic) 렌더링 파이프라인	Ours
262	103	**3**
786	309	**6**
1,835	721	**21**
3,934	1,545	**45**

표 1: 드로우 콜 호출 횟수 비교 결과(배치 용량 설정=3MB)

성능 평가는 NVIDIA RTX 4060 Laptop GPU(8GB VRAM)를 탑재한 환경에서 수행되었으며, 테스트 장면은 Sponza를 사용하였다. 평가 지표로는 (1) 측정한 FPS(Frame-Per-Second)[1], (2) 한 프레임에 호출된 드로우 콜 호출 횟수[3]를 이용하였다. 드로우 콜 횟수는 NVIDIA Nsight Graphics의 Frame Debugger로 캡처한 호출 횟수를 집계하였다. 제안하는 배치 기반 Indirect 파이프라인은 동일 장면을 단 3회, 6회, 21회, 45회의 Indirect 드로우 콜만으로 렌더링해 드로우 콜을 최대 97% 이상 절감하였음을 표 1에서 입증하였다. 단일 메모리 주소를 참고해 렌더링하므로, 기본과 제안한 방식에 있어서 렌더링 퀄리티 차이가 없다. 그림 3의 위 그래프는 그림 2의 시점 변경 전 화면에서, 삼각형 수가 증가함에 따라 FPS 변화를 나타낸 것이다. 삼각형 수가 약 15배 증가함에 따라, 기본 방식은 FPS가 급격히 하락하여 133에서 32로 크게 감소한 반면, 제안한 배치 방식은 동일한 조건에서 FPS가 163에서 119로 비교적 완만하게 감소하여 성능 감소폭을 현저히 억제하였다. 아래 그래프에서는 실시간 카메라 시점 변경에 따른 FPS 비교에서 제안한 방식은 이전보다 높은 FPS를 유지하는 반면, 기본 방식은 시점 변경 후에도 FPS가 동일한 상태를 유지하고 있다. 그림 3을 통해, 제안하는 배치 방식은 FPS가 더 높고, 변화에 대해 더 안정적인 성능을 유지하므로, 성능 향상과 성능 안정성이 동시에 개선되었음을 알 수 있습니다.

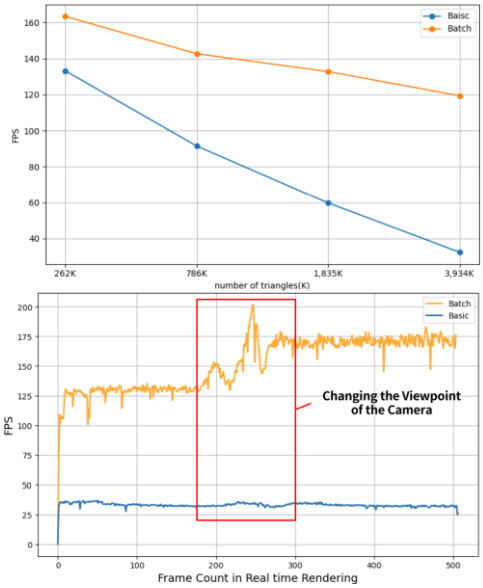


그림 3: (상)기본 및 배치 파이프라인 평균 FPS 실험 결과, (하)삼각형 3,934K 환경에서 기본 및 배치 파이프라인 실시간 FPS 실험 결과

참고문헌

[1] Matthias Wloka, "Batch, Batch, Batch:" What Does It Really Mean?, Game Developers Conference, San Jose, CA, USA, 2003. https://www.nvidia.com/docs/IO/8230/BatchBatchBatch.pdf

[2] Tristan Lorach, Vulkan: the essentials, Game Developers Conference, San Francisco, CA, USA, 2016. https://developer.download.nvidia.com/gameworks/events/GDC2016/Vulkan_Essentials_GDC16_tlorach.pdf

[3] Christoph Kubisch and Matthias Niessner, GPU Driven Rendering Pipelines, *SIGGRAPH*, Los Angeles, CA, USA, 2015.

미분 가능한 밀도 제어 기반 3D Gaussian Splatting*

김민성[01], 정문수[2], 임석현[3] 이성길[3]
[1]성균관대학교 실감미디어공학과, [2]성균관대학교 전자전기컴퓨터공학과, [3]성균관대학교 소프트웨어학과
{leon0106, moonsoo101, ish990730}@g.skku.edu, sungkil@skku.edu

Differentiable Density Control for 3D Gaussian Splatting

Minseong Kim[01], Moonsoo Jeong[2], SeokHyun Lim[3], Sungkil Lee[3]
[1]Dept. of Immersive Media Engineering, Sungkyunkwan University
[2]Dept. of Electrical and Computer Engineering, Sungkyunkwan University
[3]Dept. of Computer Science and Engineering, Sungkyunkwan University

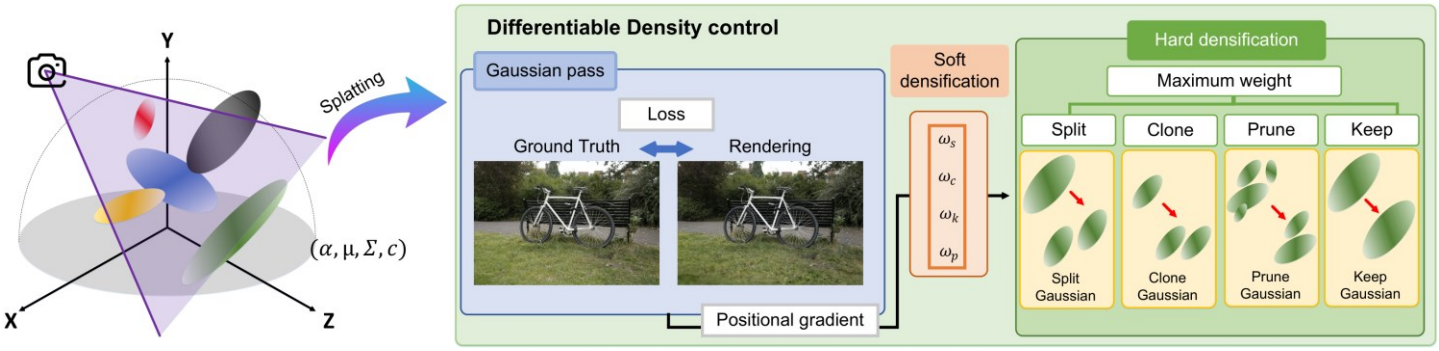


그림 1 미분 가능한 밀도 제어 기반 3DGS 개요도

요약

최근 3D Gaussian Splatting(3DGS)은 고속 렌더링과 고품질 재구성으로 주목받고 있으나, Gaussian primitive의 구조 제어에 있어 정해진 주기나 임계값 기반의 기준에 의존해 불필요한 구조 변경을 유발한다. 본 연구는 이러한 문제를 해결하기 위해 분할, 복제, 유지, 제거의 네 가지 동작에 대해 확률적 가중치를 학습하고, 가장 높은 가중치에 따라 구조를 변경하는 방식을 제안한다. 이는 기존 방식보다 유연하고 장면에 적응적인 제어를 가능하게 하여 모델의 품질과 효율을 함께 향상시킨다.

1. 서론

본 논문은 3D Gaussian Splatting(3DGS)의 구조 제어 과정에 미분 가능한 학습 기반 밀도 제어(differentiable density control)를 도입한다. 이는 경험적 기반 기존 방식보다 더 정밀한 밀도 제어가 가능하며, 높은 재구성 품질과 primitive 수를 현저히 줄이는 효과를 보인다.

* 구두(포스터) 발표논문
* 본 연구는 과학기술정보통신부 한국연구재단 중견연구자 지원 사업(RS-2024-00339681)의 지원으로 수행된 연구임
* 본 연구는 정보통신기획평가원의 메타버스 융합 대학원(IITP-2024-RS-2023-00254129), SW컴퓨팅산업원천기술개발사업(RS-2024-00454666)의 지원으로 수행된 연구임.

2. 관련 연구

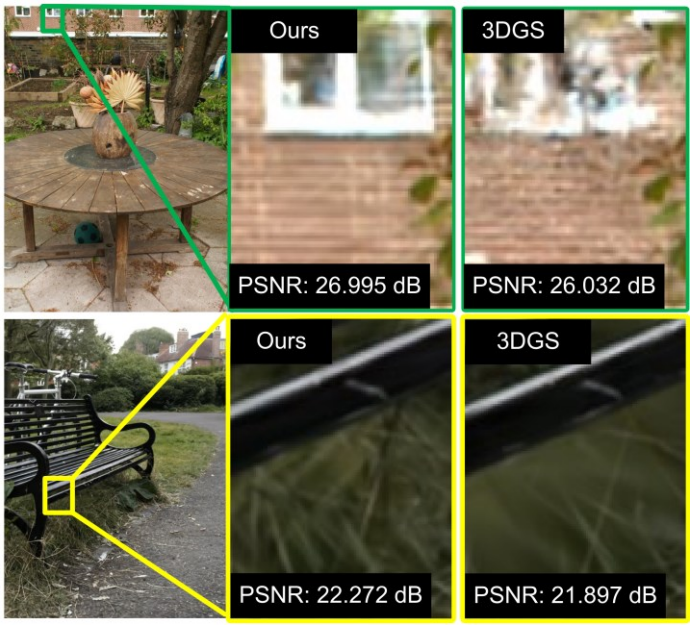
최근 3D Gaussian Splatting에서는 밀도 제어(density control)를 개선하기 위한 다양한 연구가 이루어지고 있다. AbsGS [1]와 GOF [2]는 동방향 그래디언트의 크기를 활용하여 그래디언트 상쇄 문제를 완화하고, Pixel-GS [3]는 픽셀기반 그래디언트를 도입해 픽셀 커버리지와 카메라 좌표계 깊이에 따라 Gaussian의 기여도를 조절하여 샘플링이 부족한 영역의 밀도를 선택적으로 증가시킨다.

3. 미분 가능한 밀도 제어

3.1. 실험 방법

우리는 기존 3D Gaussian Splatting(3DGS)의 밀도 제어와 다르게, 연속적으로 학습 가능한 soft densification pass를 도입하였다. 기존에는 100번째 학습마다 한 번씩 밀도 제어를 수행했으나, 본 연구에서는 이를 개선하여 처음 50번째 학습 동안은 soft densification을 통해 가중치를 학습하고, 이후 50번째 학습 동안은 hard densification을 적용하여 실제 분할, 복제, 유지, 제거의 동작을 수행하는 방식으로 밀도 제어 주기를 구성하였다. Soft densification pass에서는 분할, 복제, 유지, 제거를 위한 새로운 가중치인 w_c, w_s, w_k, w_p 가 도입된다.

그림 2 시각적 비교



이 가중치는 미분 가능하게 렌더링 손실함수로부터 역전파를 통해 업데이트되며, 학습 과정에서 손실을 최소화하는 방향으로 각 primitive에 최적화된 구조 조작을 선택하도록 유도된다. 각 가중치는 해당 primitive가 특정 동작을 수행할 확률적 중요도를 의미하며 이에 따라 기존 Gaussian primitive는 다음과 같이 가중치들 기반의 선형 결합으로 표현된다.

$$f(x;\mu,\Sigma,\omega) = w_k \cdot G(x;\mu,\Sigma) + w_s \cdot \big(G1(x;\mu,\Sigma) + G2(x;\mu,\Sigma)\big) + w_c \cdot (G1(x;\mu,\Sigma) + G2(x;\mu,\Sigma)) + w_p \cdot 0$$

밀도 제어할 Gaussian 대상을 정하는 위치 그래디언트는 AbsGS [1]의 동방향 그래디언트를 사용하였다. Soft pass에서 학습된 4개의 가중치는 다음 단계인 hard densification pass에서 최종적으로 사용된다. 구체적으로 w_c, w_s, w_k, w_p 4개의 가중치 중 가장 높은 값을 가진 항목이 해당 primitive의 최종 조작을 결정한다. 예를 들어, w_s가 가장 크다면 해당 Gaussian은 분할되어 더 세분화된 표현을 가지며, w_c 가 최대라면 동일한 Gaussian이 복제되어 주변에 추가된다. 반면, w_k가 가장 크면 현재 상태를 유지하고, w_p가 가장 큰 경우는 중요도가 낮다고 판단되어 primitive가 제거된다.

4. 결과

본 기법은 두 개의 GeForce RTX 3090 위에서 학습, 구현되었으며, 800×800의 해상도를 사용한다. Mip-NeRF 360 데이터셋의 Bicycle과 Garden 데이터를 사용하였으며, 학습 데이터셋의 매 8번째 프레임마다 테스트에 사용된다.

그림 3 Primitive 수 변화

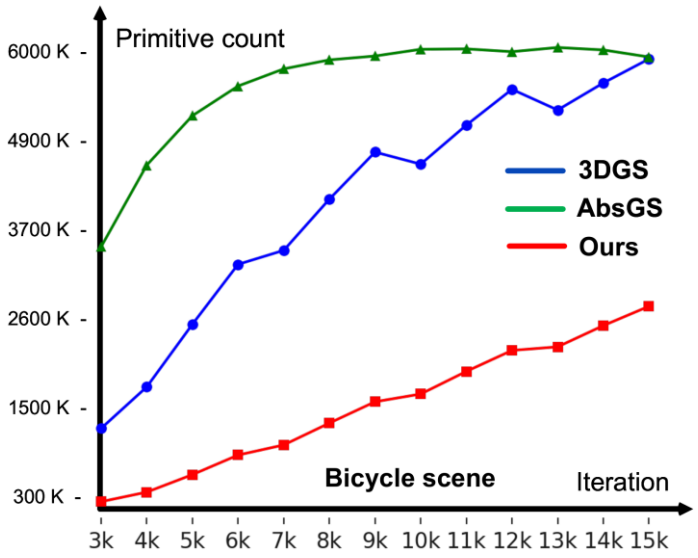


표 1 최종 primitive count

Scene	3DGS	AbsGS	Ours
Bicycle	6026K	6051K	2830K
Garden	5694K	3790K	2456K

4.1. 평가

그림 2와 3은 각각 제안한 미분가능한 밀도 제어를 적용한 렌더링 결과와 학습 과정 중의 primitive 수 변화를 나타낸다. 본 방법은 Bicycle, Garden과 같은 야외 장면에서 기존 3DGS에 비해 창문, 벤치 아래의 풀과 같은 고주파 영역을 더 정밀하게 복원하며, 구조적 세부 표현이 향상됨을 확인할 수 있다. 성능 측면에서, 동일한 장면에서의 primitive 수는 3DGS, AbsGS 대비 절반 이하로 현저히 감소하여, 메모리 효율성과 렌더링 성능 또한 크게 개선된다.

5. 결론 및 한계점

본 논문에서는 학습 가능한 가중치를 통해 장면에 적응적으로 구조를 최적화하는 미분 가능한 밀도 제어 기법을 제안하였다. 모든 가중치는 0.25로 균등 초기화하였으며, 초기값의 영향에 대한 분석은 향후 과제로 남는다.

참고문헌

[1] Ye, Zongxin, et al. "Absgs: Recovering fine details in 3d gaussian splatting." *Proceedings of the 32nd ACM International Conference on Multimedia*.
[2] Yu, Zehao, Torsten Sattler, and Andreas Geiger. "Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes." *ACM Transactions on Graphics (TOG)* 43.6 (2024): 1-13.
[3] Zhang, Zheng, et al. "Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

Visual Geometry Grounded Transformer 기반 초기화를 통한 효율적인 3D Gaussian Splatting*

김동빈[01], 이성길[1]
[1]성균관대학교 소프트웨어학과
rlaehdqls021@g.skku.edu, sungkil@skku.edu

Efficient 3D Gaussian Splatting with Visual Geometry Grounded Transformer

Dongbeen Kim[01], Sungkil Lee[1]
[1]Dept. of Computer Science and Engineering, Sungkyunkwan University

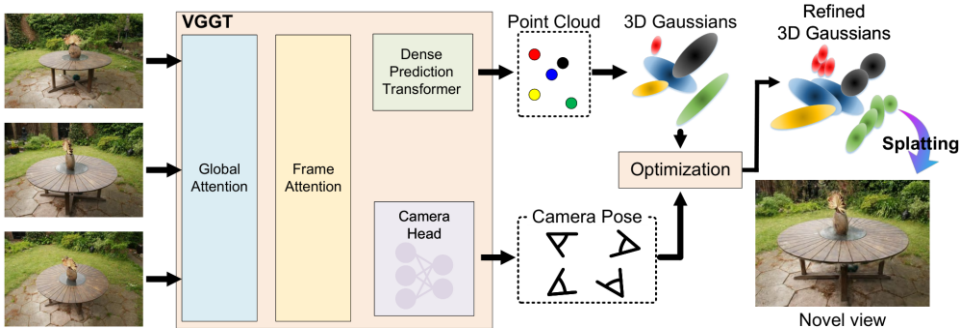


그림 1 Overview of VGGT-based Initialization for 3D Gaussian Splatting

요약

본 연구는 컴퓨터 그래픽스 분야에서 3D Gaussian Splatting(3DGS)의 초기화 개선을 탐구한다. 기존 3DGS의 COLMAP을 통한 초기 포인트 클라우드와 카메라 포즈를 추정하는 방식에서 벗어나, 최근 제안된 Visual Geometry Grounded Transformer(VGGT)를 활용한다. 이를 통해 빠르고 정확하게 카메라 포즈와 초기 포인트 클라우드를 추정하여 더욱 효율적으로 신규 시점에서 고품질의 이미지를 합성하는 것을 목표로 하며, 결과적으로 기존 3DGS보다 높은 정확도를 달성하였다.

1. 서론

3DGS는 고품질의 새로운 시점 합성을 위한 효과적인 기술로 자리 잡았다. 그러나 기존의 3DGS 방법은 초기 포인트 클라우드와 카메라 포즈 추정 단계에서 COLMAP에 의존하고 있어 많은 계산 비용이 소요되며, 이미지 수가 적거나 시점 간 중복이 부족할 경우 정확도가 떨어진다. 본 연구에서는 이러한 문제를 해결하기 위해 VGGT를 활용하여 보다 정확한 초기 데이터를 제공함으로써 3DGS 성능 향상에 대한 방안을 제안한다. 또한, 일반적인 조건뿐만 아니라 sparse-view 조건에서의 실험도 수행하여 VGGT와 3DGS 간의 연계 가능성을 실증적으로 검토하며, 적용가능성에 대해 최초로 탐색한다.

* 구두(포스터) 발표논문
* 본 연구는 과학기술정보통신부 한국연구재단 중견연구자 지원사업(RS-2024-00339681), 정보통신기획평가원 SW컴퓨팅산업원천기술개발사업(RS-2024-00454666)의 지원으로 수행된 연구임.

2. 관련 연구

Kerbl et al. [1]은 3D Gaussian을 활용한 Splatting 기법으로 빠른 학습과 실시간 렌더링을 가능하게 하였다. 3DGS는 초기 포인트와 카메라 포즈 추정에 Schönberger et al. [2]의 COLMAP 기반 Structure-from-Motion(SfM) 방식을 활용한다.

Wang et al. [3]은 Transformer 기반의 기하학 예측 모델을 통해 이미지를 입력으로 받아 카메라 포즈와 3D 포인트를 예측하는 방식을 제안하였다. VGGT는 sparse view에서도 강건한 성능을 보여주며, 기존 SfM 기반 초기화의 대안으로 주목받고 있다.

3. 3DGS를 위한 VGGT 기반 초기화

3.1. VGGT를 통한 Point Cloud 및 Pose 추정

본 연구에서는 VGGT를 활용하여 이미지의 공간적 관

표 1 Quantitative results under full-view supervision

	PSNR↑	SSIM↑	LPIPS↓
3DGS	25.955 dB	0.877	0.155
VGGT-GS	**27.916 dB**	**0.900**	**0.124**

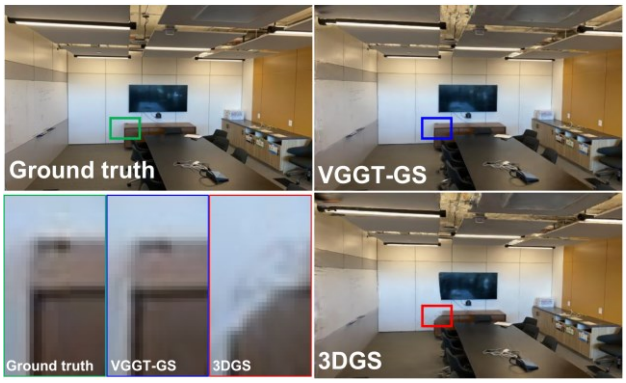


그림 2 Qualitative Comparison under full-view supervision

계와 기하학적 특성을 효과적으로 추출하여 정확한 카메라 포즈와 초기 포인트 클라우드를 생성함으로써 빠르고 정확한 초기화를 가능하게 한다.

3.2. VGGT 초기화를 사용한 3DGS 최적화

VGGT로 생성된 초기 포인트 클라우드와 카메라 포즈를 이용하여 기존의 3DGS 최적화 과정을 수행한다. 보다 정확하고 안정적인 초기값을 활용함으로써 최적화 단계의 수렴 속도와 안정성이 크게 향상된다.

4. 결과

본 기법은 한 개의 Intel(R) Core(TM) i9-10900 2.8GHz, GeForce RTX 3090 위에서 학습, 구현되었으며, 이미지는 518×310의 해상도를 사용한다. 전체 이미지 중에서 매 8번째 이미지를 테스트용으로 할당하고, 나머지 이미지는 학습용으로 사용하였다.

4.1. 전체 View 환경에서의 성능 비교

그림 2와 표 1은 전체 뷰를 활용해 초기화 및 최적화를 수행한 결과를 보여준다. VGGT-GS는 구조 보존과 디테일 표현에서 우수하며, 정량적 지표에서도 3DGS 대비 모든 지표에서 일관된 성능 향상을 보인다.

4.2. Sparse View 환경에서의 성능 비교

그림 3과 표 2는 24장의 이미지만 사용한 결과로, VGGT-GS는 적은 iteration만으로도 3DGS보다 나은 성능을 기록했다. 이는 초기 품질과 최적화 효율성 측면에서 VGGT-GS의 강점을 잘 보여준다.

4.3. View 수에 따른 초기화 성공률 분석

표 3은 다양한 view 수와 데이터셋에서의 초기화 성공률을 비교한 것이다. VGGT-GS는 모든 조건에서 안정적으로 초기화에 성공한 반면, 3DGS는 적은 view 조건에서 실패가 잦았다. 이는 VGGT-GS의 강건한 초기화 성능을 보여준다.

표 2 Quantitative results under 24-view at 7000 iteration

	PSNR↑	SSIM↑	LPIPS↓
3DGS	25.785 dB	0.747	0.266
VGGT-GS	**27.318 dB**	**0.808**	**0.219**

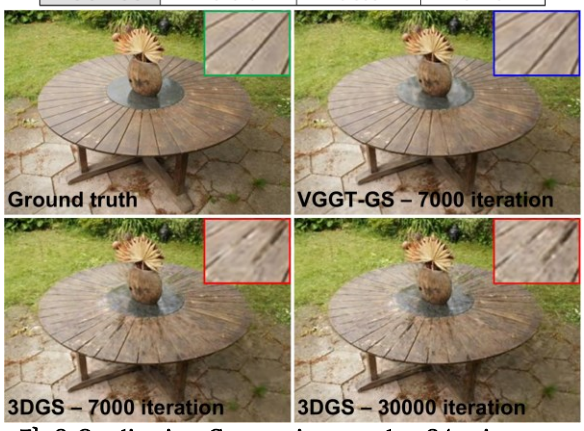


그림 3 Qualitative Comparison under 24-view setup

표 3 Dataset 및 View 수에 따른 모델 별 초기화 성공률

Dataset	LLFF		Mip-NeRF360		
Model \ View	3-view	Full-view	3-view	24-view	Full-view
3DGS	37.5%	100%	33%	66%	100%
VGGT-GS	100%	100%	100%	100%	100%

5. 결론

본 논문에서는 VGGT를 3DGS의 초기화 과정에 성공적으로 적용하여 기존 COLMAP 방식의 주요 한계점을 효과적으로 개선하였다. VGGT의 빠르고 정확한 초기화는 전체적인 효율성과 재구성 품질을 향상시켰으며, 소수의 뷰 환경에서도 높은 초기화 성공률을 기록하였다.

향후 연구로는 이미지 Warping과 Diffusion 기술을 활용하여 소수의 뷰 환경에서 더욱 효과적인 신규 시점 합성을 할 수 있는 연구를 진행할 예정이다.

참고문헌

[1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis.3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023.
[2] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
[3] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

가상 패딩을 통한 밉맵 기반 이미지 인터폴레이션 최적화*

전수연[01], 박재인[2], 손주희[2], 이성길[2]
[1]성균관대학교 실감미디어공학과, [2]성균관대학교 소프트웨어학과
{archblossom, jaynap, juheeson}@g.skku.edu, sungkil@skku.edu

Virtual Padding for Optimized Mipmap-Based Image Interpolation

Suyeon Jeon[01], Jaein Park[2], Juhee Son[2], Sungkil Lee[2]
[1]Dept. of Immersive Media Engineering, Sungkyunkwan University
[2]Dept. of Computer Science and Engineering, Sungkyunkwan University

그림 1: 가상 패딩을 통한 밉맵 기반의 이미지 인터폴레이션 파이프라인과 메모리 사용량 비교

요약

본 연구는 GPU 기반 다중 해상도 이미지 인터폴레이션(interpolation)에서, 2의 제곱수가 아닌 (Non-Power-of-Two, NPOT) 해상도의 입력 이미지에 존재하는 결손 영역을 메모리 효율적으로 복원하기 위한 기법을 제안한다. 제안된 방법은 NPOT 해상도 이미지에 대해 실제 메모리 할당 없이 가상 패딩을 적용하여 다중 해상도 처리를 수행하고, 인터폴레이션 결과를 별도의 후처리 없이 원본 해상도에 직접 생성한다. 해당 구조는 기존 방식과 동일한 품질의 인터폴레이션 결과를 유지하면서, GPU 메모리 사용량을 줄이는 결과를 보인다.

1. 서론

이미지 인터폴레이션을 통한 결손 영역 복원인 인페인팅(inpainting)은, 결손 영역의 크기가 클수록 단일 해상도에서의 인터폴레이션만으로는 구조적 일관성을 유지하기 어려워 경계 왜곡이나 시각적 부자연스러움이 발생할 수 있다. 이를 해결하기 위해, 전역 구조의 추정과 세부 정보 복원을 동시에 수행할 수 있는 다중 해상도 기반의 이미지 피라미드 접근 방식이 효과적으로 활용된다. 특히, 이미지 피라미드의 밉맵을 사용한 풀-푸쉬(Pull-Push) 알고리즘 [1]은 저해상도 방향으로 진행되는 분석 단계인 풀과 고해상도 방향으로 진행되는 합성 단계인 푸쉬를 통해 결손 영역을 점진적으로 인터폴레이션하는 구조로 널리 사용된다. 그러나 밉맵은 이미지 해상도가 2의 제곱수 (Power-of-Two, POT)가 아닐 경우, 가장 가까운 상위 POT 해상도로 패딩해야 하며, 고해상도에서는 불필요한 메모리 낭비가 커지는 문제가 있다. 본 연구는 이러한 한계를 해결하기 위해, NPOT 이미지를 실제 메모리 패딩 없이 POT처럼 처리할 수 있는 가상 패딩 구조를 제안하고, 이를 풀-푸쉬 알고리즘에 효과적으로 통합한 GPU 기반 인페인팅 기법을 구현한다.

2. 관련 연구

2.1. 이미지 피라미드

Burt et al. [2]은 이미지의 공간적 스케일 정보를 계층적으로 표현하기 위한 이미지 피라미드 구조를 제안하였다. 이 구조는 원본 이미지를 반복적으로 절반 해상도로 다운샘플링하여 다중 해상도 계층을 구성하며, 인터폴레이션, 필터링의 다양한 이미지 처리 작업에 널리 활용된다.

* 구두(포스터) 발표논문
* 본 연구는 과학기술정보통신부 한국연구재단 중견연구자 지원사업(RS-2024-00339681)의 지원으로 수행된 연구임.
* 본 연구는 정보통신기획평가원의 메타버스 융합 대학원(IITP-2024-RS-2023-00254129), SW컴퓨팅산업원천기술개발사업(RS-2024-00454666)의 지원으로 수행된 연구임.

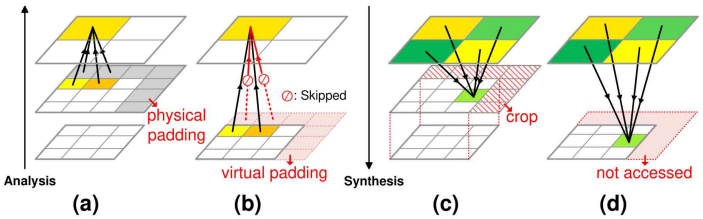


그림 2: 기존 방식 (a), (c)의 물리적 패딩과 제안 방식 (b), (d) 의 가상 패딩을 적용한 분석 및 합성 과정 비교.

밉맵은 Williams [3]가 제안한 이미지 피라미드 기반 텍스처 필터링 기법으로, 앨리어싱 완화와 성능 향상을 위해 각 레벨이 POT 해상도로 구성된다. 원본 이미지가 POT 해상도가 아닐 경우, 밉맵 생성을 위해 패딩이 필요하며 이로 인해 메모리 낭비가 발생할 수 있다.

2.2. 풀-푸쉬 기반 결손 영역 인터폴레이션

Strengert et al. [1]는 알파 채널을 활용해 결손 영역에 대한 유효성 마스킹을 수행하고, 평균 기반 피라미드 방식의 반복적인 다운샘플링을 통해 밉맵을 생성하는 풀 단계 (분석)와, 이차 B-스플라인 가중치를 기반으로 상위 밉 레벨의 정보를 참조하여 인터폴레이션하는 푸쉬 단계 (합성)를 결합한 풀-푸쉬 기반 인페인팅 기법을 제안하였다. 각 단계는 인접한 밉 레벨을 참조함으로써 결손 영역을 점진적으로 복원한다.

3. 가상 패딩을 활용한 밉맵 기반 풀-푸쉬

본 연구는 NPOT 해상도 이미지에 대해 Strengert et al. [1]의 풀-푸쉬 구조를 유지하면서도, 메모리 및 연산 효율 향상을 위해 두 가지 구조적 개선을 제안한다. 기존 방식은 상위 POT 해상도를 위한 물리적 패딩과 원본 해상도로의 크롭 과정을 필요로 했으나, 본 연구는 가상 패딩을 적용하여 실제 메모리 할당 없이 밉맵 생성을 시작하고, 크롭 과정 없이 결과 이미지를 직접 복원함으로써 불필요한 메모리 사용과 연산을 줄인다.

3.1. 가상 패딩 기반 밉맵 생성 최적화

밉맵 생성을 위한 첫 번째 밉 레벨 텍스처는 원본 NPOT 텍스처에 가상 패딩을 적용한 풀 연산의 결과이다. 가상 패딩은 GPU 연산에서 텍스처가 상위 POT 해상도를 갖는 것처럼 접근하되, 패딩 영역의 텍셀은 평균 계산에서 제외되도록 처리된다. 이에 따라 유효한 2×2 텍셀 블록만을 기반으로 평균값을 계산하여 텍스처를 생성하며, 이는 기존 방식에서 두 번째 밉 레벨에 해당한다. 그러나 본 연구에서는 이를 밉맵 생성의 초기 레벨로 간주함으로써 하나의 레벨 생성을 생략하고, 메모리 사용과 연산량을 절감한다.

3.2. 크롭 과정을 생략한 가상 패딩 기반 인페인팅

본 연구는 가상 패딩 기법을 적용하여 상위 POT 해상도의 텍스처 생성을 생략하므로, 해당 해상도로의 푸쉬 연산 과정이 불필요하다. 이에 따라 상위 POT 해상도에서 원본 해상도 외의 영역에 대한 크롭 과정 없이, 하위 POT 해상도에서 푸쉬 연산을 한 번 더 수행하는 것만으로 동일한 품질의 결과를 직접 생성할 수 있다.

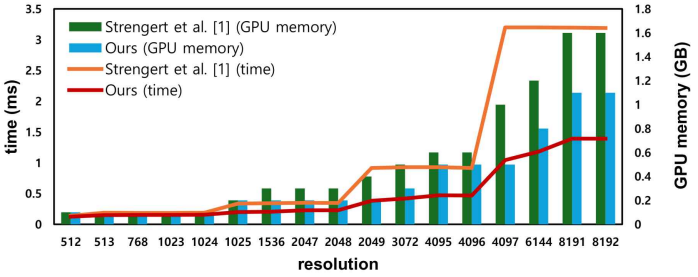


그림 3: 이미지 해상도별 인페인팅 속도와 GPU 메모리 사용량 비교 결과

Hole 영역 비율 (%)	4	16	36	64	81
Strengert et al. [1] (ms)	0.343	0.346	0.350	0.351	0.352
Ours (ms)	0.198	0.200	0.203	0.205	0.207

표 1: 결손 영역 크기에 따른 인페인팅 속도 비교 결과

4. 결과

4.1. 측정 방법

성능 평가는 Intel Core i9-10900 CPU와 NVIDIA RTX 2080 GPU 환경에서 수행되었다. 다양한 해상도의 이미지에 동일한 위치와 크기의 결손 영역을 설정하여 Strengert et al. [1] 방식과 비교하였으며, PSNR과 SSIM을 통해 두 기법의 결과 이미지 품질이 동일함을 확인한 후, 프레임당 평균 렌더링 속도와 GPU 메모리 사용량을 측정하였다.

4.2. 측정 결과

그림 3과 표 1은 본 연구와 Strengert et al. [1]의 인페인팅 성능을 각각 해상도와 결손 영역 비율에 따라 비교한 결과이며, 그림 3의 막대그래프는 GPU 메모리 사용량을 함께 나타낸다. 실험 결과, 원본 해상도와 상위 POT 해상도 간의 비율이 0.5에 가까울수록 성능 차이가 뚜렷하게 나타났으며, 해상도가 높아질수록 본 연구의 기법은 더 적은 메모리와 연산으로 빠른 처리 속도를 보였다. 또한, 모든 결손 영역 비율에서 제안 기법은 기존 방식 대비 일관되게 낮은 인페인팅 시간을 기록하였다.

참고문헌

[1] Markus Strengert, Martin Kraus, and Thomas Ertl. 2006. "Pyramid Methods in GPU-based Image Processing." In Proceedings of Vision, Modeling, and Visualization 2006, 169–176.

[2] Peter J. Burt. 1981. "Fast Filter Transform for Image Processing." Computer Graphics and Image Processing 16, 1 (1981), 20–51.

[3] Lance Williams. 1983. "Pyramidal Parametrics." In Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '83), 1–11.

Interaction with Virtual Objects using Human Pose and Shape Estimation

Hong Son Nguyen[1] DaEun Cheong[1○] Andrew Chalmers[2] Myoung Gon Kim[1]
Taehyun Rhee[3] JungHyun Han[1*]

[1]Korea University, [2]Victoria University of Wellington, [3]The University of Melbourne

nguyenhongson303@gmail.com, {wjdekdms001, m_gon_kim, jhan}@korea.ac.kr
andrew.chalmers@vuw.ac.nz, taehyun.rhee@unimelb.edu.au

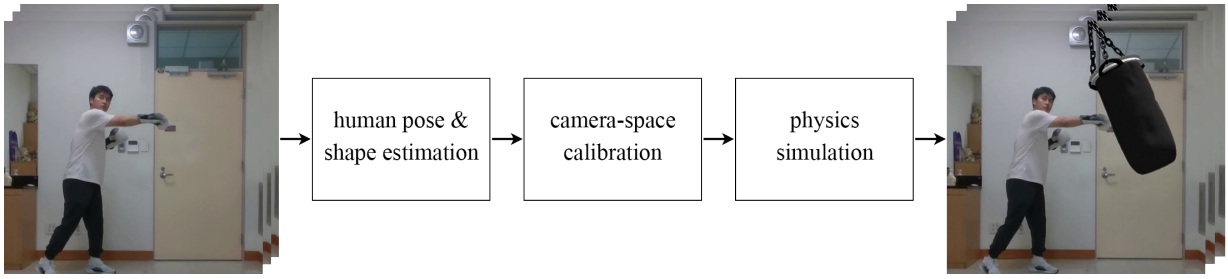


Figure 1: Our AR system is composed of three modules that facilitate interactions between a real human and virtual objects. In this AR boxing game, the user is interacting with the virtual punching bag.

Abstract

In this paper, we propose an AR system that facilitates a user's natural interaction with virtual objects in an augmented reality environment. The system consists of three modules: human pose and shape estimation, camera-space calibration, and physics simulation. The first module estimates a user's 3D pose and shape from a single RGB video stream, thereby reducing the system setup cost and broadening potential applications. The camera-space calibration module estimates the user's camera-space position to align the user with the input RGB image. The physics simulation enables seamless and physically natural interaction with virtual objects. Two prototyping applications built upon the system prove an enhancement in the quality of interaction, fostering a more immersive and intuitive user experience.

Keywords: Augmented reality, Full-body interaction, Pose and shape estimation.

1 Introduction

For various augmented reality (AR) applications, it is essential to provide natural and seamless interactions between real humans and virtual objects. To achieve this, recent methods estimate 3D human pose and shape directly from RGB inputs, thereby avoiding the need for bulky sensors and enabling collision-aware virtual scenarios. In this paper, we propose an AR system that takes an RGB video as input, estimates a user's 3D pose and shape, and facilitates the user's interaction with 3D virtual objects. As shown in Figure 1, the system comprises three modules for (1) human pose and shape estimation, (2) camera-space calibration, and (3) physics simulation.

2 Methodology

In our AR setup, a user moves on the floor and the large screen in front of the user displays the mirrored view of the environment using an RGB webcam, which is mounted on top of the screen and captures the *full body* of the user. The captured environment is augmented with virtual objects so that the user can interact with them.

Among the three modules presented in Figure 1, this paper focuses on the second module, *camera-space calibration*, while briefly sketching the first and third modules, for which there exist many off-the-shelf methods and commercial/open-source programs. We employ HMR [1] to estimate SMPL-X [2] parameters

(θ, β) and derive the 3D mesh $\mathcal{V}$ and joints $\mathcal{J}$.

In our setup, the human model is invisible to the user, but its mesh is used, for example, to detect *collisions* with the virtual objects. In the current implementation, the physics simulation module is implemented using Unity engine. However, it can be seamlessly replaced by other commercial or open-source engines.

Previous HMR employs the *weak-perspective camera model*, which is implemented as an *orthographic projection* followed by a *scaling*. Consequently, the reconstructed human model is not correctly located in the 3D coordinate system of the real camera, e.g., its camera-space depth may be over- or under-estimated. Then, not only would we miss the *collision* between the human model and virtual objects, but we would also handle the *occlusion* between them incorrectly. Therefore, the human model must be *calibrated* so that it is transformed into the *full-perspective* camera space.

The essential step for calibration is to estimate the camera-space coordinates of the *root joint*. Once they are estimated correctly, all the other joints' coordinates can be determined immediately using their positions relative to the root, which are stored in the 3D pose $\mathcal{J}$.

For this, we employed a deep learning network proposed by Pavllo *et al.* [3], which processes "a sequence of 2D poses" to output the camera-space positions of the root. It was modified and retrained to take "a single frame of 2D pose" as input and estimate the root's camera-space position.

Let $\mathcal{M}$ denote the network and j denote the 2D pose input to $\mathcal{M}$. By projecting the 3D pose $\mathcal{J}$ onto the image plane, we can obtain j. It is simple to implement because HMR returns not only θ and β but also the parameters of the weak-perspective camera model, i.e., the scale factor $s \in \mathbb{R}$ and the 2D translation vector $t \in \mathbb{R}^2$. The 2D pose j is obtained using the 3D pose $\mathcal{J}$ as follows:

$$j = s\Pi(\mathcal{J}) + t \tag{1}$$

where $\Pi(\cdot)$ denotes orthographic projection. $\mathcal{M}(j)$ is the root's coordinates, (x, y, z), in the camera space.

The calibration network, $\mathcal{M}$, is trained using the Human3.6M dataset [4]. Note that the Human3.6M dataset contains only "square" images with aspect ratio one. In contrast, our system presented in Figure 1 is designed to take general "rectangular" images. Consequently, the x- and y-coordinates estimated by $\mathcal{M}$ are not reliable.

Table 1: Calibration model evaluation.

# input frames	parameters	FLOPs	trajectory errors
1 frame	**4.24M**	**8.46M**	**147.0mm**
27 frames	8.51M	16.99M	147.7mm
81 frames	12.70M	25.38M	181.3mm

As the Human3.6M dataset lacks rectangular input support, we use only the z-coordinate from $\mathcal{M}(j)$ and compute (X, Y) using the root pixel location (u, v) and camera intrinsics $\mathbf{K}$ via standard back-projection:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = z\mathbf{K}^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \tag{2}$$

3 Discussion & Conclusion

This paper presents a practical solution for enabling natural interactions between humans and 3D virtual objects in AR environments using only a monocular RGB camera. Built upon efficient and accurate deep learning techniques, our system reconstructs and calibrates the human model to robustly manage collision and occlusion.

In general, HPS methods based on weak-perspective projection yield inaccurate 3D pose estimates, especially when the subject is near the camera. However, in our AR setup (**??**), where the user is intentionally positioned at a fairly long distance, the weak-perspective assumption becomes more valid, enabling effective body reconstruction.

Unlike prior trajectory estimation methods that are computationally heavy and do not estimate 3D body shape, we leverage 2D poses derived from the estimated 3D joints $\mathcal{J}$ to condition a modified version of Pavllo *et al.* [3]'s temporal network. Our calibration model, operating on single-frame input, achieves lower trajectory error and reduced FLOPs compared to models using 27 or 81 input frames (see Table 1), possibly due to avoiding irrelevant past-frame accumulation.

Despite these strengths, our approach inherits certain limitations. The use of SMPL limits precise 3D localization, particularly for extreme or uncommon poses such as lying down. As our trajectory model is trained on standard datasets, it may generalize poorly to such cases. This constrains its applicability to relatively upright motions typical of gaming and entertainment contexts.

To expand the applicability of our method to more diverse scenarios, future work will explore advanced trajectory conditioning, improved body representations, and generalization to broader motion types, thus enhancing the overall quality of AR interaction.

Acknowledgements

This research was supported by the Ministry of Science and ICT, Korea, under the ITRC (Information Technology Research Center) Support Program (IITP-2024-2020-0-01460).

References

[1] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.

[2] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[3] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7753–7762.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

AutoVRTest: 가상현실 환경에서 AI 에이전트를 활용한 맵 탐색 및 공간 오류 검출 자동화 기술 개발[1)]

유승한 나민수 김동환 김성기*
조선대학교 컴퓨터공학과
qewr08@chosun.ac.kr, drone1575@chosun.ac.kr, cuactpg@chosun.ac.kr, skkim@chosun.ac.kr

AutoVRTest: Development of Automated Map Exploration and Spatial Error Detection Technology Using AI Agents in Virtual Reality Environments

Yu SeungHan Na MinSoo Kim DongHwan SeongKi Kim
Dept. of Computer Science, Chosun University

요약

본 논문은 Unity 기반 가상현실(VR, Virtual Reality) 콘텐츠에서 사용하는 지도의 자동화된 테스트를 위한 강화학습 기반 에이전트를 설계하고 구현하였다. 본 연구에서 수작업 테스트의 한계를 보완하기 위해, 해당 에이전트는 Proximal Policy Optimization(PPO)과 Curiosity-Driven exploration 기법을 활용해 VR 환경을 자율적으로 탐색하고 오류 가능 구간을 식별할 수 있었다. 픽셀 기반 입력과 VR 컨트롤러 시뮬레이션을 통해 인간과 유사한 상호작용을 수행하며, 일반적인 이동으로는 접근이 어려운 영역까지 테스트가 가능하다. 이 방식은 테스트 효율성과 커버리지를 향상시키며, 다양한 VR 응용 분야로의 확장 가능성도 지니고 있다.

키워드: 게임 테스팅, AI, VR, 강화학습, 환경 탐색
Keywords: Game Testing, AI, VR, Reinforcement Learning, Environment Exploring

1. 서론

최근 Unity 기반 가상현실 콘텐츠가 교육, 의료, 시뮬레이션, 게임 등 다양한 분야에서 활발히 제작되고 있다. 이에 따라 가상 공간의 품질 보증(QA, Quality Assurance)의 중요성도 점차 부각되고 있다. 특히 VR 콘텐츠는 사용자의 몰입도와 직접적인 상호작용에 기반하기 때문에, 맵 내에 존재하는 물리적 오류, 상호작용 불가 오브젝트, 또는 시야 블라인드 구간 등은 사용자 경험에 큰 영향을 미친다.
그러나 현재 대부분의 VR 맵 테스트는 수작업에 의존하고 있으며, 이로 인해 테스트 시간의 증가, 반복 작업에 따른 오류 누락, 그리고 테스트 커버리지의 불균형 문제가 지속적으로 제기되고 있다. 복잡한 상호작용 요소를 포함하는 VR 콘텐츠의 특성상 이러한 문제는 더욱 두드러지며, 개발 과정의 병목 지점으로 작용하고 있다.
또한, 현재 Unity 엔진은 NavMesh 기반의 AI 경로 탐색 시스템을 제공하고 있으나, 단일 목적지에만 이동이 가능하며, 맵 전체를 자율 탐색하거나 미 탐색 영역을 인식하여 스스로 이동하는 기능은 제공되지 않는 상황에 있다.
본 연구에서는 이러한 한계를 극복하고자, Unity 기반 VR 콘텐츠에서 맵을 자동으로 탐색하며 오류 가능성을 검출하는 테스트 에이전트를 설계하는 것을 목표로 한다. 제안하는 에이전트는 사람처럼 VR 환경을 둘러보며, 상호작용이 가능한 객체들과 물리적 행동을 수행함으로써, 인간 테스터가 발견하기 어려운 영역까지도 효과적으로 탐색할 수 있도록 하였다.

2. 관련 연구

게임 테스트 자동화는 반복적이고 비용이 많이 드는 수작업을 대체하기 위한 기술로 주목받고 있으며, 최근에는 강화학습(DRL, Deep Reinforcement Learning)을 활용한 연구가 활발하다. 본 절에서는 관련 연구를 세 가지 흐름으로 정리한다.

2.1 픽셀 기반 DRL 에이전트

픽셀 기반 DRL은 게임 화면 이미지만으로 학습하는 방식으로, DQN, OpenAI Five, VizDoom 등이 대표 사례이다. 이 방식은 내부 정보 없이 시각 입력만으로 학습이 가능해 실제 게임 환경에 적용이 용이하다. 본 연구 또한 Unity 환경의 1인칭 시점 화면을 입력으로 사용한다는 점에서 유사한 접근을 취한다.

2.2 사용자 선호 기반 DRL

최근에는 명시적 보상 대신 사용자 선호 데이터를 반영하여 에이전트를 학습시키는 방식도 등장하고 있다.[1] 예를 들어, Preference-conditioned 에이전트는 사람이 선호할 경로를 학습함으로써 인간 친화적인 테스트를 수행한다. 본 연구는 이를 직접 활용하지는 않지만, 향후 인간 유사성 강화를 위한 참고 사례로 활용 가능하다.

2.3 Unity 기반 자동화 도구

Unity 환경을 기반으로 한 자동화 연구로는 UI 상호작용, 시나리오 반복 등을 자동화한 연구의 사례[2]가 있다. 그러나 해당 연구는 정적 테스트 중심이며, 본 연구와 같이 환경을 자율적으로 탐색하는 VR 기반 에이전트와는 차별성이 존재한다.

---

* 구두(포스터) 발표논문
* 본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음(2024-0-00062)
* 교신저자: 저자 이름 옆에 * 표시
1) 학부생 주저자 논문임

3. 연구 설계

그림 1: 해당 연구에서 테스트로 사용할 탐색 환경

본 연구에서는 그림 1과 같은 Unity 기반 가상현실 콘텐츠 내의 맵을 자율적으로 탐색하고 오류 가능 구간을 검출할 수 있는 테스트 에이전트를 설계하였다. 이를 위한 에이전트는 강화학습 기법을 기반으로 구성되며, 특히 Proximal Policy Optimization(PPO) 알고리즘과 Curiosity-Driven Exploration 기법을 결합하여, 사전 정의된 목표 없이도 미탐색 영역을 중심으로 자율적인 맵 탐색을 수행할 수 있도록 설계된다.
에이전트는 Unity 환경에서 VR 컨트롤러의 동작을 시뮬레이션할 수 있도록 구성된다. 즉, 실제 사용자처럼 양손 컨트롤러를 이용하여 오브젝트를 잡거나 하는 등 물리적인 상호작용이 가능하다. 이를 위해 Unity 엔진의 물리 기반 상호작용 시스템과 연동되어, 충돌 판정, 잡기(grab)를 감지할 수 있는 인터페이스를 포함한다.
탐색 방식은 다음과 같은 절차로 구성된다:

환경 초기화 및 상태 수집: Unity 환경 내에서 에이전트는 현재 위치, 시야, 주변 오브젝트 정보를 픽셀 단위로 수집하며, 시야 내의 상호작용 가능 대상과 장애물 여부를 판단한다. 이 정보는 에이전트의 상태(state)로 인코딩된다.
탐색 정책 결정: PPO 알고리즘은 수집된 상태 정보를 기반으로 이동 방향, 시선 회전, 손의 움직임 등 다음 행동을 결정한다. curiosity 모듈은 새로운 지역이나 상태 변화가 큰 행동에 보상을 부여함으로써, 에이전트가 반복적으로 동일한 경로를 탐색하는 것을 방지한다.
물리 상호작용 수행: 에이전트는 양손 컨트롤러로 물체를 집거나 문을 여는 등 물리 상호작용을 수행하며, 해당 과정에서 충돌 오류, 비정상적인 반응, 혹은 상호작용 실패 등의 이벤트를 감지한다. 이 과정은 오류 가능성 탐지에 핵심적인 역할을 한다. 또한 컨트롤러의 물리적 상호작용으로 일반적으로 가기 힘든 위치의 탐색을 유도하도록 한다.

그림 2: 탐색 어려운 구역

탐색 불가능 영역 대응: 물리 상호작용에서 언급하였듯이 에이전트는 일반적인 이동으로는 접근이 어려운 지역, 예를 들어 협소한 공간이나 특정 동작(점프, 잡기 등)을 통해서만 도달 가능한 구역까지 탐색하려 시도하며, 이러한 영역에서 오류 발생 확률이 높은 것으로 가정한다.
전체 시스템은 강화학습 기반의 자율성, VR 인터페이스의 상호작용성, 그리고 Unity 엔진의 실시간 처리 기능을 통합적으로 활용하여, VR 콘텐츠의 사전 테스트 효율을 크게 향상시키는 것을 목표로 하였다.

4. 결론

본 연구는 기존 VR 콘텐츠 제작 환경에서 간과되기 쉬운 테스트 자동화의 가능성을 제시하고, 콘텐츠 품질 향상을 위한 새로운 접근을 탐색하였다. 기존의 수작업 QA 방식은 테스트 범위가 제한적이고 반복 작업에 많은 인력이 소요되는 구조적 한계를 지닌다. 이에 본 연구에서는 강화학습 기반의 테스트 에이전트를 설계하여, 자율 탐색 능력을 통해 사람의 개입 없이도 VR 맵 전반을 능동적으로 탐색하고, 상호작용 가능한 오브젝트에 대해 실제 사용자와 유사한 행동을 모사하여 테스트를 수행하는 접근법을 제안하였다.
이러한 자동화된 테스트 방식은 기존의 수작업 대비 시간과 비용 면에서 효율성을 개선할 수 있을 뿐만 아니라, 테스트 커버리지 측면에서도 인간 테스터가 놓치기 쉬운 구역이나 상호작용을 보다 효과적으로 포괄할 수 있을 것으로 기대된다. 또한, VR 콘텐츠의 특수성인 양손 상호작용 및 공간 제약 조건을 고려한 설계를 통해, VR 환경에 특화된 테스트 자동화 기법 구현의 기술적 가능성을 제시하였다.
다만, 본 연구는 현재 설계 및 구현이 완료된 초기 단계에 있으며, 정량적 성능 평가나 시연 결과, 정성적 사례 자료 등은 추후 실험을 통해 확보될 예정이다. 향후에는 에이전트의 탐색 로그, 오류 검출 사례, 시각화된 경로 및 탐색 커버리지 등 정성·정량적 데이터를 체계적으로 수집·분석하고, 이를 시각 자료로 구성하여 연구의 신뢰성과 실효성을 강화할 계획이다. 또한 실제 사용자 행동 데이터를 학습에 반영하거나, 오류 발생 가능성이 높은 영역을 우선 탐색하는 정책을 도입함으로써 더욱 정밀하고 사용자 맞춤형 QA 자동화 도구로의 발전도 모색할 예정이다.
제안한 방법론은 게임뿐만 아니라 교육용 VR, 의료 시뮬레이션, 산업 훈련 콘텐츠 등 다양한 응용 분야에 적용 가능성이 있으며, 향후 실증적 연구와 성능 평가를 통해 확장성과 실용성을 더욱 구체화할 수 있을 것으로 기대된다.

참고문헌

[1] S. Abdelfattah, A. Brown and P. Zhang, "Preference-conditioned Pixel-based AI Agent For Game Testing,"2023 IEEE Conference on Games (CoG), Boston, MA, USA, 2023, pp. 1-8, doi: 10.1109/CoG57401.2023.10333200.
[2] S. Park, D. Kim, and W. Lee, "UnityPGTA: A Unity Platformer Game Testing Automation Tool Using Reinforcement Learning" *Journal of KIISE*, vol. 51, no. 2, pp. 149-156, 2024. doi: https://doi.org/10.5626/JOK.2024.51.2.149
[3] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," inIEEE Transactions on Neural Networks, vol. 9, no. 5, pp. 1054-1054, Sept. 1998, doi: 10.1109/TNN.1998.712192.
[4] D. Pathak, P. Agrawal, A. A. Efros and T. Darrell, "Curiosity-Driven Exploration by Self-Supervised Prediction,"2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 488-489, doi: 10.1109/CVPRW.2017.70.

XR Interaction Toolkit을 이용한 VR 드로잉 시스템 구현

용상임[0], 윤종현, 김선정
한림대학교 소프트웨어학부 콘텐츠IT 전공
{sangim04113, yjh07271}@gmail.com, sunkim@hallym.ac.kr

Implementation of a VR Drawing System Using the XR Interaction Toolkit

Sang-im Yong[0], Jong-Hyun Youn, Sun-Jeong Kim
Major of Contents IT, Division of Software, Hallym University

요약

유니티 6 엔진과 XR Interaction Toolkit을 활용하여 VR 공간에서 드로잉이 가능한 시스템을 구현하였다. 사용자는 해당 공간에서 펜을 잡고 그림을 그릴 수 있으며, UI를 통해 컬러피커, 펜 굵기 조절, 색상 미리보기 등의 기능을 사용할 수 있다. 또한 선 스타일과 레이어 기능을 지원하여 다양한 표현이 가능하도록 설계되었다.

1. 서론

VR 기술의 발전은 예술과 기술의 융합을 촉진하고 있으며, 몰입형 3D 환경에서 창작과 전시가 가능한 플랫폼이 제공되고 있다. 특히 VR 드로잉은 3차원 공간에서 사용자가 직접 그림을 그릴 수 있게 하여 전통적 매체의 한계를 극복하고 새로운 창작의 가능성을 제시한다. 대표적인 사례로는 Google "Tilt Brush"가 있으며, 이는 VR 기반 창작 도구로서의 가능성을 입증한 바 있다.

본 시스템은 유니티 6의 XR Interaction Toolkit을 기반으로 드로잉 환경을 구현하였으며, VR 드로잉의 몰입감과 표현력을 지원하기 위한 기능적 요소에 중점을 두었다.

2. 시스템 설계

본 시스템은 사용자가 최소한의 학습만으로도 VR 환경에서 드로잉을 수행할 수 있도록 XRGrabInteractable을 활용하여 펜 오브젝트와 컨트롤러 간의 상호작용을 구성하였다. 사용자가 다양한 창의적 표현을 할 수 있도록 컬러 휠 기반의 UI, 선 마감 스타일, 텍스처 애니메이션 등 세부 기능을 정교하게 설계하였다. 특히 UI는 사용자가 간단한 조작만으로 원하는 색상과 펜 굵기를 직관적으로 조정할 수 있도록 구성하였다. 또한, 사용자가 작업을 효율적으로 수행할 수 있도록 레이어 기반의 드로잉 방식을 적용하였다.

* 학부생 주저자 논문 / 포스터 발표논문
* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아서 수행된 연구임(No. RS-2023-00254695).
* 본 연구는 2025년도 과학기술정보통신 및 정보통신기획평가원의 'SW중심대학사업' 지원을 받아 수행되었음(2024-0-00064)

3. 시스템 구현

3.1. 펜 상호작용과 드로잉 오브젝트 생성 과정

사용자가 VR 공간에서 펜 오브젝트를 잡아 드로잉할 수 있도록 XRGrabInteractable 컴포넌트를 적용하였으며, 컨트롤러와 오브젝트 간 충돌 및 트리거 입력을 감지하여 펜을 잡고 놓는 동작을 자동으로 인식한다.

펜을 잡은 상태에서 트리거 입력이 감지되면, 실시간 드로잉이 시작된다. 새로운 GameObject가 동적으로 생성되고, LineRenderer 컴포넌트가 부착되며, 첫 번째 정점 위치는 펜 끝 위치로 설정된다. 이후 매 프레임마다 펜 끝 위치를 추적하면서, 이전 정점과의 거리가 0.01m 이상일 경우 새로운 정점이 추가되어 선이 자연스럽게 이어지도록 구현하였다.

트리거 입력이 해제되면 드로잉이 종료되고, 이후 다시 선을 그릴 경우 새로운 LineRenderer가 생성된다.

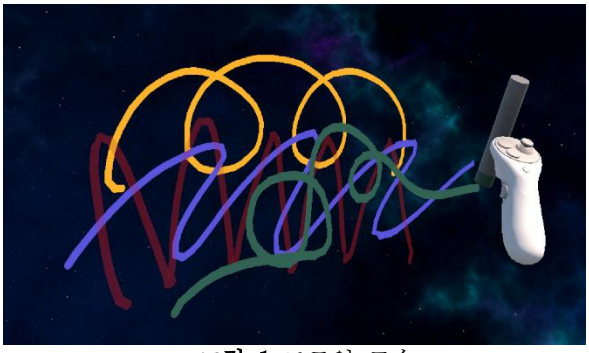


그림 1 드로잉 모습

3.2. UI를 통한 색상 및 펜 굵기 조절

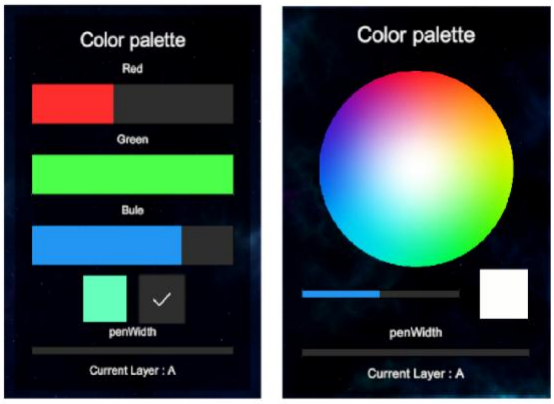
초기 구현 단계에서는 RGB 슬라이더를 통해 각각의 RGB 값을 개별적으로 조절하여 색상을 설정했으나, 직관성과 시각적 편의성을 높이기 위해 컬러 휠 기반 UI로 개선하였다.

개선된 컬러피커는 VR 컨트롤러의 Ray가 컬러 휠 이미지 위에 투사되고, 충돌 지점의 픽셀 색상을 추출하는 방식으로 구현되었다. 사용자는 컬러 휠에서 원하는 위치를 가르키고 트리거를 눌러 간편하게 색상을 선택할 수 있으며, 선택한 색상은 즉시 펜에 반영된다.

또한, 명도 조절 기능이 포함된 슬라이더를 추가하여

색상 선택의 폭을 확장하였다. 명도 조절 슬라이더 옆에는 컬러 미리보기 이미지 기능이 있으며 선택한 색상이 시각적 피드백으로 제공된다. 펜 굵기 조절 슬라이더는 사용자가 값을 변경하면 penWidth 값이 업데이트되고, 이는 드로잉 시 생성되는 LineRenderer의 폭에 반영되도록 설계되었다.

이처럼 UI는 사용자의 직관적인 인터랙션과 창작 편의성을 중심으로 설계되었으며, 단순한 조작만으로도 다양한 색상과 선 굵기를 자유롭게 조절할 수 있는 환경을 제공한다. 이러한 구성은 VR 드로잉의 몰입감을 높이는 데 기여한다



개선 전 개선 후

그림 2 UI 개선 전과 개선 후

3.3. 선 마감 스타일

다양한 선 마감 표현을 위해 AnimationCurve를 활용하여 시간에 따라 선의 두께를 조절하였다. 특히 선의 양 끝을 부드럽게 둥글리거나 날카롭게 마무리하는 방식을 개별적으로 설계함으로써, 실제 필기도구(예: 연필, 붓)의 시각적 특성을 효과적으로 재현하였다.

부드러운 선 마감은 LineRenderer.numCapVertices 속성을 통해 구현되며, 이 값은 선의 양 끝을 반원 형태로 보간하는 정점 수를 의미한다. 또한 꺾이는 지점에서는 numCornerVertices 속성을 조정하여 선이 자연스럽게 곡선 형태로 이어지도록 하였다.

뾰족한 선 마감은 LineRenderer.widthCurve에 사용자 정의 AnimationCurve를 적용해 구현된다. 이 곡선은 시작과 끝에서 너비가 0에 가까워지고, 중간 구간에서 최대 두께를 유지하는 형태로 구성되어 실제 브러시나 펜촉의 흐름을 시각적으로 잘 표현한다. 이러한 기법은 VR 드로잉 환경에서 현실의 필기 도구와 유사한 표현을 구현하는 데 효과적으로 활용될 수 있다.

그림 3 선 마감 스타일

3.4. 선 텍스처 애니메이션

점선, 패턴, 무지개 색상 등 다양한 텍스처가 선을 따라 움직이는 듯한 시각 효과를 주기 위해, material.mainTextureOffset 속성이 시간에 따라 동적으로 변경되도록 설계하였다. Update() 함수 내부에서 Time.deltaTime에 비례하여 오프셋 값을 증가시킴으로써, 텍스처가 선을 따라 흐르는 애니메이션이 구현된다. 이 방식은 드로잉 동작과 텍스처 움직임을 연동시켜 보다 생동감 있는 시각 효과를 제공하며, 사용자 몰입도와 표현의 완성도를 동시에 높이는 데 기여한다.

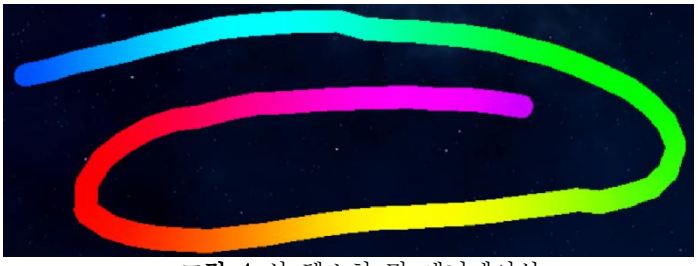


그림 4 선 텍스처 및 애니메이션

3.5. 레이어 추가 및 관리

사용자가 선을 그릴 때마다 시스템은 자동으로 새로운 GameObject를 생성하고, 이를 지정된 레이어에 배치한다. 이러한 구조는 각 LineRenderer를 독립적으로 관리할 수 있게 하며, 사용자가 원하는 시점에 해당 레이어의 표시 여부를 손쉽게 제어할 수 있게 한다. 이를 통해 복잡한 드로잉 구성도 체계적으로 관리할 수 있는 기반을 제공한다.

4. 향후 계획

현재 텍스처 애니메이션은 mainTextureOffset 조정에 기반하고 있으나, 향후 UV 좌표를 직접 수정하여 렌더링 성능을 향상시킬 계획이다.

또한 유니티 Shader Graph를 활용하여 빛나는 선 효과나 벽 너머에서도 보이는 특수 효과 등 고급 시각 표현 기법을 적용함으로써, 표현력을 강화하고자 한다.

레이어 기능 역시 확장하여 사용자가 드로잉한 오브젝트를 자유롭게 이동, 회전 및 복사할 수 있도록 지원함으로써 복잡한 작업도 직관적으로 구성할 수 있는 환경을 제공할 예정이다.

이러한 개선을 통해 일반 사용자에게도 향상된 사용 경험과 창작 효율성을 제공하며, 궁극적으로는 VR 드로잉 시스템을 창의성, 몰입감, 표현력을 두루 갖춘 창작 도구로 발전시키는 것을 목표로 한다.

참고문헌

[1] L. Ramsier, “Evaluating the Usability and User Experience of a Virtual Reality Painting Application,” 석사학위논문, University of North Carolina at Chapel Hill, 2019.

[2] 윤희선, 정진헌, 뉴미디어 공공미술의 확장성 연구 – VR 드로잉을 중심으로, 디지털융복합연구, 2021.

[3] R. Rodriguez, B. T. Sullivan, “An Artist’s Perspectives on Natural Interactions for Virtual Reality 3D Sketching,” Proceedings of the ACM CHI Conference on Human Factors in Computing Systems, 2024.

NeRF에서의 SIFT 기반 광선 할당*

최영준[0], 최준서, 정승화[1]
세종대학교 소프트웨어학과
{topjuney, wnstj010126}@gmail.com, seunghwajeong@sejong.ac.kr

SIFT-Guided Ray Allocation for Neural Radiance Fields

Young-Jun Choi[0], Jun-Seo Choi, Seung-Hwa Jeong[1]
Dept. of Software, Sejong University

요약

본 연구는 NeRF(Neural Radiance Field)에서 동일한 연산 비용 하에 더 정밀한 장면 복원을 위한 학습 데이터 샘플링 방법을 제안한다. SIFT(Scale Invariant Feature Transform)기반 비균등 확률 분포 맵을 통해 NeRF 학습에 높은 기여를 하는 데이터를 우선 샘플링하여 학습 초기부터 높은 정확도를 제공한다. 이미지 엣지를 활용한 NeRF 학습 데이터의 샘플링 방법과 정확도 비교 연구를 통해, 제안한 방식의 효과성을 입증하였다.

1. 서론

NeRF[1]는 복잡한 장면의 3차원 구조와 시점에 따른 조명을 하나의 연속적인 함수로 표현하여 새로운 시점에서의 이미지를 합성하는 기술로, 학습 데이터는 무작위 샘플링 방식을 이용하여 Ray(광선)로 추출한다. 학습 속도를 개선하기 위해 광선 샘플링 단계에서 확률 맵을 이용하는 연구들이 진행되고 있다. 대표적인 연구인 EGRA-NeRF(Edge-Guided Ray Allocation NeRF)[2]는 모든 이미지에 대해 Canny 윤곽선 검출을 활용해 엣지 맵을 생성하고, 확률 맵을 이용해 비균등 광선 샘플링을 수행한다. 본 논문에서는 3D 관점에서 의미가 있는 일관된 시점의 매칭점을 제공하는 SIFT를 이용하여 NeRF 학습 샘플링을 위한 기존 연구를 개선하는 방법을 제안한다.

2. 관련 연구

2.1. Neural Radiance Fields

NeRF[1]는 새로운 시점의 이미지를 합성하기 위해 3차원 공간상의 위치(x,y,z)와 해당 위치를 바라보는 방향(θ,φ)을 입력 받아 해당하는 점의 색(c)과 밀도(σ)를 예측하는 5차원 연속 함수 $F_\theta:(x,y,z,\theta,\varphi)\rightarrow(c,\sigma)$를 학습한다. 이 학습과정은 카메라 중심에서 특정 픽셀을 향해 나아가는 3D 직선인 Ray(광선) 단위로 이루어지며, 각 Iteration마다 훈련 이미지들로부터 무작위 샘플링 된 픽셀들에 대응하는 광선의 batch가 구성된다. 이러한 균등 Ray 샘플링(uniform ray sampling)은 단순하고 구현이 쉽지만, 정보 밀도가 낮은 영역에서의 비효율적인 계산이 이루어지고, 고주파수(high-frequency)정보가 많은 장면의 표현에 미흡해지는 한계가 있다.

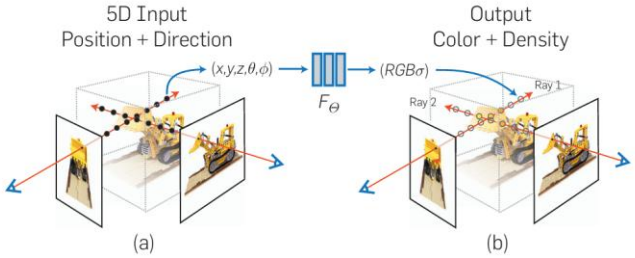


그림 1: NeRF[1]의 학습 과정

2.2. Edge-Guided Ray Allocation

EGRA-NeRF[2]는 NeRF의 균등 Ray 샘플링의 한계를 극복하기 위해 제안된 방식으로, 윤곽선 영역에 Ray를 집중시켜 더 정밀한 장면 복원을 유도한다. 입력 이미지에 Canny 엣지 검출을 적용해 엣지 픽셀에는 높은 샘플링 확률을, 그 외 픽셀에는 낮은 확률을 부여하여 비균등 확률 분포 $P(x,y)$를 만들고, 확률 분포에 따라 픽셀을 샘플링 한다. 이후 과정은 기본 NeRF와 동일하게 진행된다. 실제로 PSNR, SSIM, LPIPS 성능 지표에서 원본 NeRF 대비 일관된 성능 향상을 보인다. 하지만 EGRA-NeRF는 2D 입력 이미지에서의 윤곽선만을 이용하기 때문에 의미 없는 텍스처의 경계가 강조되거나 노이즈에 따라 엣지가 잘못 검출될 수 있다는 한계점을 가지고 있다.

* 포스터 논문
* 학부생 주저자 논문임

3. 방법: SIFT-Guided Ray Allocation(SIFT-NeRF)

SIFT[3]는 이미지에서 크기(scale)나 회전에 불변한 특징점(keypoint)을 추출할 수 있어, 널리 사용되는 대표적인 특징점 검출 알고리즘이다. NeRF의 Ray 샘플링 단계에서 SIFT 확률 맵을 구성하는 것은 엣지 기반 확률 맵에 비해 다음과 같은 장점을 가진다. SIFT 특징점은 스케일 및 회전에 불변하고, 여러 시점 이미지의 동일한 3D 지점에서 안정적으로 검출되기 때문에, 시점 변화에 따라 위치나 크게 형상이 바뀌기 쉬운 Canny 엣지에 비해 더 일관된 시점간 대응점을 제공한다. SIFT는 Gaussian scale-space에서 극댓값을 검출하므로 엣지 검출에 비해 노이즈에 강하다. 또한 Canny 엣지 검출은 2D 이미지에서 텍스처 패턴, 노이즈 등 불필요한 픽셀까지 다수 샘플링 되지만 SIFT를 이용한다면 [그림 2]과 같이 3D 기하 복원에 핵심적인 영역에 집중 샘플링을 수행한다. 이를 통해, NeRF 학습에 필요한 데이터를 효율적으로 배분하여 학습을 가속화할 수 있다.

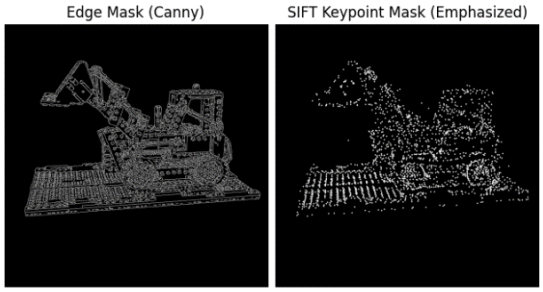


그림 2: Canny 엣지 검출 결과와 SIFT 특징점 검출 결과

본 연구에서는 모든 훈련 이미지에 대해 특징점을 찾아내고, 특징점이 아닌 영역도 일정 확률로 학습되도록 특징점 픽셀에는 높은 샘플링 확률을, 특징점 외의 픽셀에는 낮은 확률을 부여하는 EGRA-NeRF와 동일한 확률 분포 생성 방식을 사용하였다. 다만 Canny에서 SIFT를 적용하는 방식으로 변경하면서 검출된 픽셀의 수가 현저히 줄어들기 때문에 EGRA-NeRF에서 특징점 픽셀에 부여했던 샘플링 확률보다 SIFT 특징점에 더 높은 샘플링 확률을 부여하였다.

4. 실험 및 결과

본 연구에서는 한정된 시간 내 학습 정확도 향상을 확인하기 위해 최신 중요도(엣지) 기반 학습 데이터 샘플링 방식인 EGRA-NeRF[2]와 정확도 비교 연구를 진행하였다. 모든 방식의 학습 시간은 10k iteration 당 약 12분이 소요되었으며, 각 iteration에 동일하게 1,024개의 픽셀 샘플을 할당하였다. 샘플링을 위해 100장의 입력 이미지에 Canny 엣지 검출과 SIFT를 적용한 결과, 각각 0.328초 및 6.591초가 추가로 소요되었으나, 학습 시간에 비해 지연 영향은 거의 무시할 수 있는 수준이었다. [그림 3]은 NeRF[1], EGRA-NeRF[2]와 비교한 정량적인 품질 향상을 보여준다. 동일 iteration 대비 정확도(PSNR) 향상을 확인하였다. 또한, [그림 4]는 기존 연구와 비교한 시각적 차이를 보여주며, 기존 연구 대비 엣지 주변이 더 선명해지는 것을 확인할 수 있다.

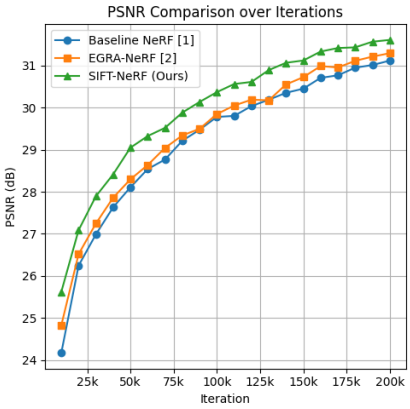


그림 3: NeRF[1], EGRA-NeRF[2]와 SIFT-NeRF의 PSNR

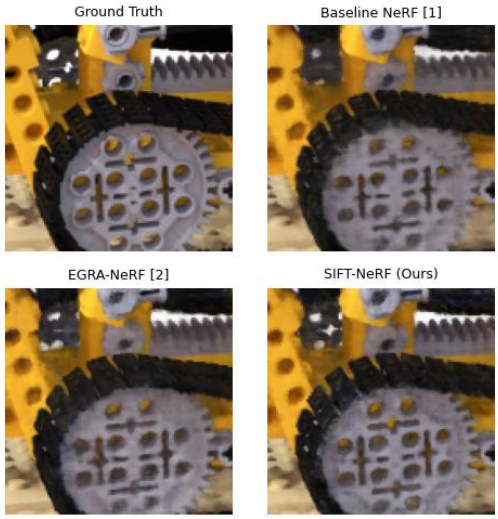


그림 4: Ground Truth와 생성 이미지 비교. SIFT-NeRF는 경계와 세부 묘사에서 향상된 품질을 보여줌

4. 결론

본 연구의 목표는 NeRF의 학습과정 중 픽셀 샘플링 단계에서 SIFT 기반 확률 맵을 활용하여 3D 공간상의 중요 특징점 위주로 Ray를 배분함으로써 한정된 시간 내에서 효과적으로 시각 품질을 향상시키는 것이다. 그 결과, 기존 NeRF 및 EGRA-NeRF 대비 PSNR 향상이라는 정량적 성능 개선을 달성하였다. EGRA-NeRF가 2D 엣지 중심의 Ray 분포를 사용하는 것과 달리, 본 연구는 SIFT 기반의 3D 특징점을 활용한 Ray 분포 전략을 통해, 정보량이 높은 영역을 효과적으로 학습함으로써 시각 품질 개선을 보여주었다.

참고문헌

[1] Mildenhall, B. et al., “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” ECCV, 2020.

[2] Gai, Z., Liu, Z., Tan, M., Ding, J., Yu, J., Tong, M., & Yuan, J. (2023). EGRA-NeRF: Edge-Guided Ray Allocation for Neural Radiance Fields. Image and Vision Computing, 134, 104670.
[3] Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110.

3D Gaussian Splatting 뷰어에서의 초해상화 적용에 대한 연구*

조희석[0], 이제희[0], 최준서, 정승화[1]
세종대학교 소프트웨어학과
{joyrock0611, jeh6778, wnstj010126} @gmail.com, seunghwajeong@sejong.ac.kr

A Study on the Application of Super-Resolution in 3D Gaussian-Splatting Viewer

Hee-seok Cho[0], Jeh-hui Lee[0], Jun-seo Choi, Seung-hwa Jeong[1]
Dept. of Software, Sejong University

요약

본 연구는 3DGS (3D Gaussian Splatting)의 렌더링 뷰에 초해상화 기술 (Super-Resolution, SR)을 적용하여 품질을 향상시키는 방법을 제안한다. 사용자 시점에서 SR을 진행함으로써 렌더링 시간은 기존 3DGS의 것을 유지하고, 사용자가 보는 영역의 품질을 향상시킨다. MSE 및 GAN 기반의 SR 모델을 3DGS에 적용하였으며 정량적 품질 향상을 확인하였다. 또한, 이를 이용한 연구에 대한 방향성을 제시한다.

1. 서론

3DGS (3D Gaussian Splatting) [1]의 품질을 개선하기 위해 SR (Super-Resolution)을 적용하는 연구가 활발히 이루어지고 있다. 이는 SR 모델 및 GS 파라미터 학습의 연산 부담을 늘려 GS가 가진 빠른 학습 속도의 이점이 상쇄된다.

본 연구에서는 이러한 연산적 부담을 줄이기 위해 SR의 적용 시점을 렌더링 이후, 보이는 시점에서만 뷰어 단계에서 적용하는 방법을 제안한다. 3D GS가 원본 해상도로 렌더링하면 뷰어 단에서 SR을 적용하여 세부 디테일을 복원한다. 이러한 구조를 통해 3D GS의 고품질 결과물을 얻을 수 있다.

본 연구에서는 3DGS의 결과물과 두 개의 SR 모델을 이용하여 GS 렌더링 이후 SR을 적용한다. 모델은 MSE 기반 모델(EDSR[2]), GAN 기반 모델(ESRGAN[3])을 활용하였으며, 정량적, 시각적 결과물을 바탕으로 향후 연구 방향성을 제시한다.

2. 관련 연구

2.1. 3D Gaussian-Splatting

[0] 공동1저자
[1] 교신저자
* 포스터발표논문
* 학부생 주저자 논문임

[그림 1] 전체 파이프라인

3D Gaussian-Splatting 은 공간을 다수의 3차원 가우시안(Gaussian)을 생성하여 렌더링하는 기법이다 [1]. SfM (Structure-from-Motion) [2]을 통해 얻은 포인트 클라우드를 기반으로 각 점마다 초기 3D Gaussian을 생성한다. 각 Gaussian은 중심 위치, 공분산, 색상, 투명도 파라미터를 가진다. 이들은 카메라 파라미터에 따라 이미지 평면상에 2D Gaussian 형태로 투영되고, Rasterizer를 통해 이미지가 구성된다. 마지막으로 렌더링 된 이미지와 Ground Truth 이미지 간의 오차를 기반으로 Loss가 계산되며, 이 손실로부터 발생하는 Gradient는 역전파를 통해 각 파라미터를 업데이트한다.

2.2. Super-Resolution

SR 기술은 한 장의 저해상도 이미지(LR)을 입력받아 이미지를 고해상도로 만드는 기술로, 초기엔 SRCNN [3], EDSR [4], VDSR [5]등 CNN 기반 회귀 모델이 PSNR 중심의 발전을 이뤘다. 이후 SRGAN, ESRGAN [6] 등 생성형 모델 위주의 발전이 이루어져 왔다.

3. 실험

3.1. 구현

본 연구는 [그림 1] 과 같은 파이프라인으로 구성되어 렌더링(Rendering side)은 3DGS로 생성된 원본 이미지를 사용하고 사용자에게 직접 보이는 뷰어 단계(Viewer side)에서 SR을 통해 세부 정보를 복원한다. 렌더링 단계에서 직접적으로 SR을 적용하지 않고 보이는 영역만 SR을 적용하였다. 3DGS의 학습은 NeRF Synthetic Dataset을 이용하여 진행하였다. SR 모델은 DIV2K 데이터 셋을 이용하여 학습을 진행하였다.

[그림 2] 왼쪽부터 Render View, Ground Truth, 3DGS를 적용한 원본, 3DGS 결과물에 EDSR 및 ESRGAN을 적용한 결과

[표 1] 모델별 정량적 분석 결과

	3DGS 적용 원본	3DGS [1] + EDSR [4]	3DGS [1] + ESRGAN [6]
PSNR	30.10	29.69	31.49
SSIM	0.9386	0.7960	0.9532
LPIPS	0.1258	0.1346	0.0439

3.2. 결과

[그림 2]는 3DGS의 렌더링 결과에 각각의 모델을 적용하여 비교한 이미지이다. EDSR보다 ESRGAN이 PSNR 상으로 더욱 선명하게 세부 디테일을 복원하였다. [표 1]은 본 연구의 실험의 결과를 정량적 수치(PSNR, SSIM, LPIPS)로 나타낸 표이다. 지표 측정은 GS 학습 과정에서 데이터셋의 해상도를 1/4로 다운스케일링하여 렌더링을 한 뒤, SR을 적용하여 원본 해상도로 복원하여 진행하였다.

4. 논의 및 향후 연구

[그림 2]와 같이 ESRGAN이 EDSR보다 시각적인 품질 향상을 시각적으로 확인되며 [표 1]과 같이 LPIPS의 수치는 확연히 좋아졌으나, 원본 이미지의 구조가 변형됨을 확인하였다. 이는, MSE 손실을 결합하거나 주파수 분리 [8]를 통해 원본의 구조를 유지하면서 고주파수 영역에만 GAN을 적용함으로써, 원본의 구조를 유지할 수 있을 것이다.

EDSR 적용 시에는 [그림 2]와 같이 LR 이미지보다 약간 선명해졌으나 [표 1]과 같이 정량적 정확도 지표에서는 낮은 결과를 보여주었다. 이는 Bicubic 커널 기준의 이미지 데이터셋을 이용한 학습 방식으로써 3D GS 렌더링 방식과 상이하기 때문에 효과적인 적용이 어려웠다. 또한, 실시간 렌더링이 불가하였으며, 경량화 모델을 적용하기에 품질 향상을 기대하기 어렵다. 이에 향후 연구 방향은 SR 모델의 씬 특화 학습 및 경량화를 통해 렌더링의 실시간성을 달성하면서, 품질을 개선할 수 있는 연구를 진행할 것이다. NEALS [9]에서 소개된 실시간 CNN 학습 및 경량화 방법을 본 연구에 활용하면 추가 실시간 렌더링 및 GS 특성에 특화된 SR 모델 생성이 가능할 것으로 믿는다.

5. 결론

본 연구는 3DGS의 Viewer side에서 SR을 적용하여 영상의 세부정보를 복원하여 품질을 향상시키는 방법을 제안한다. MSE기반 모델과 GAN 기반 모델을 통해 업스케일링을 진행하며 어떤 모델의 성능이 더 효과적으로 작용하는지에 대해 실험하였으며, PSNR 및 LPIPS의 증가를 정량적으로 확인하였다. 특히, 최근 진행되고 있는 GS와 SR의 결합의 관점에서 입력 이미지의 SR 적용이 아닌 Viewer side에서 SR을 진행하는 것이 효과를 가질 수 있다는 것을 보임을 통해 GS와 SR 결합에 대한 파이프라인에 새로운 가능성을 제시한다.

Reference

[1] KERBL, Bernhard, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023, 42.4: 139:1-139:14.

[2] SCHONBERGER, Johannes L.; FRAHM, Jan-Michael. Structure-from-motion revisited. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 4104-4113.

[3] DONG, Chao, et al. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 2015, 38.2: 295-307.

[4] LIM, Bee, et al. Enhanced deep residual networks for single image super-resolution. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017. p. 136-144.

[5] KIM, Jiwon; LEE, Jung Kwon; LEE, Kyoung Mu. Accurate image super-resolution using very deep convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 1646-1654.

[6] WANG, Xintao, et al. Esrgan: Enhanced super-resolution generative adversarial networks. In: *Proceedings of the European conference on computer vision (ECCV) workshops*. 2018. p. 0-0.

[7] YU, Xiqian, et al. Gaussiansr: 3d gaussian super-resolution with 2d diffusion priors. *arXiv preprint arXiv:2406.10111*, 2024.

[8] FRITSCHE, Manuel; GU, Shuhang; TIMOFTE, Radu. Frequency separation for real-world super-resolution. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019. p. 3599-3608.

[9] JEONG, Seunghwa, et al. Real-time CNN training and compression for neural-enhanced adaptive live streaming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

회전 원판을 활용한 고정 카메라 기반 SfM-free Gaussian Splatting 학습 방법

김규민[0], 이도해[1], 백하늘[1], 이인권[1]
연세대학교 지능융합협동과정[0], 연세대학교 컴퓨터과학과[1]
{kkm8121, dlehgo1414, bbhn2024, iklee}@yonsei.ac.kr

A Fixed-Camera-Based SfM-free Gaussian Splatting Training Utilizing a Rotating Turntable

Kyu Min Kim[0], Dohae Lee[1], Hanul Baek[1], In-Kwon Lee[1]
Dept. of Intelligence Convergence[0], Dept. of Computer Science[1], Yonsei University

요약

본 논문은 고정된 카메라로 회전 원판 위의 물체를 촬영한 이미지를 이용하여 Structure-from-Motion (SfM) 과정 없이 3D Gaussian Splatting을 학습하는 새로운 방법을 제안한다. 회전 원판 환경에서는 배경이 고정되고 물체만 회전하므로 SfM 과정 전 배경 제거 전처리가 필수적이다. 그러나 이로 인해 특징점이 부족해져 기존 SfM 기반 기법이 자주 실패하는 문제가 발생한다. 이를 해결하기 위해, 제안하는 방법은 회전 원판을 사용해 촬영했다는 사전 지식을 바탕으로, Gaussian Splatting 최적화 과정에서 가우시안 속성과 함께 원판의 회전축 및 회전각을 동시에 추정한다. 이를 통해 SfM 과정을 거치지 않고도 정확한 3D 복원이 가능해진다. 실제 및 가상 데이터를 대상으로 한 실험 결과, 제안하는 방법은 SfM 없이 Gaussian Splatting을 최적화하는 기존의 방법 대비 더욱 높은 성능을 보였다.

1. 서론

이미지를 기반으로 3D 객체를 재구성하는 기술은 가상현실, 게임 등 다양한 응용 분야에 활용 가능성이 높아 활발히 연구되고 있다. 최근에는 짧은 학습 시간과 실시간 렌더링이 가능한 3D Gaussian Splatting(3DGS)[1]이 주목받고 있다. 3DGS는 대부분의 이미지 기반 3D 객체 복원 기법처럼 Structure-from-Motion(SfM)[2]을 통해 카메라 변수를 추정하는 전처리 과정이 필요하다. 한편, 3DGS 기술을 활용하여 실물 객체를 3D 재구성하려는 상황에서, 회전하는 원판 위에 물체를 놓고 고정된 카메라로 촬영하는 방식은 공간적 제약 등의 문제를 효과적으로 해결할 수 있다는 이점이 있다. 이 경우 배경은 고정된 상태에서 물체만 회전하므로 SfM을 통해 물체와 카메라의 상대적 위치를 올바르게 추정하기 위해 배경제거 전처리가 필수적이다. 그러나, 이로 인해 이미지의 특징점 수가 부족해져 SfM을 통해 카메라 변수를 올바르게 추정하지 못하는 문제가 발생한다. 이를 해결하기 위해 COLMAP-Free 3D Gaussian Splatting(CF-3DGS)[3], SfM-free 3D Gaussian Splatting(SfM-free GS)[4]과 같은 SfM을 필요로 하지 않는 3DGS 학습 방법 등이 제시되었지만, 본 연구와 같이 배경 제거로 인해 특징점이 제한된 상황에는 여전히 잘 작동하지 않는 한계가 있다. 본 논문에서는 물체가 회전 원판 위에서 회전한다는 사전 지식을 활용하여 3DGS 최적화 과정에서 가우시안의 속성뿐만 아니라 원판의 회전축과 회전각을 동시에 추정하는 새로운 기법을 제안한다. 실제 및 가상 데이터를 활용한 실험을 통해, 제안한 기법이 회전 원판을 활용해 촬영한 이미지셋으로부터 기존 기법 대비 더 정확하게 3D 객체를 재구성함을 보였다.

2. 방법

2.1. 3D 가우시안 중심축 회전 기반 렌더링

기존의 3DGS는 이미지마다 카메라 변수가 다르지만, 본 연구에서는 카메라 변수는 고정되어 있고, 대신 이미지에 따라 물체의 회전각이 달라진다고 가정한다(그림 1(a)). 제안하는 방법은 SfM을 사용하지 않기 때문에 가우시안 집합의 초기 상태를 무작위 포인트 클라우드로 설정하고, 초기 회전축은 카메라의 up 벡터 방향으로 설정한다. 원판의 회전각은 일정한 각속도를 가정하여 초기화하고, 실제 회전의 미세한 오차를 보정하기 위해 작은 MLP 네트워크 기반 잔차 예측 모듈을 도입한다(그림 1(b)). 학습 과정에서는 추정된 회전축과 회전각을 기준으로 가우시안 집합을 회전시킨 후(그림 1(d)), 이를 렌더링하여 얻은 이미지와 ground truth(GT) 이미지 간의 손실을 계산한다. 이 손실값을 바탕으로 가우시안 속성 뿐만 아니라 회전각의 잔차를 예측하는 MLP 모듈과 회전축을 동시에 최적화한다.

* 포스터 발표논문
* 본 논문은 요약논문(Extended Abstract)로 연구의 초기결과임.
* 이 연구는 정부 (과학기술정보통신부)의 재원으로 한국연구재단 (No. RS-2024-00348094) 및 한국전파진흥협회 (No. RNIX20230200)의 지원으로 수행되었음.

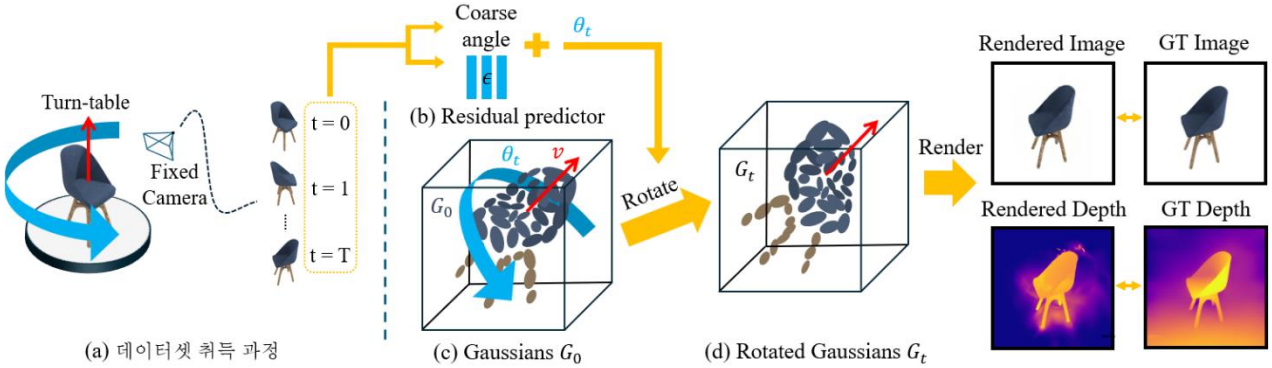


그림 1: 전체 파이프라인, (a) 고정된 카메라로 회전하는 원판 위의 물체를 촬영하여 최대 timestep T까지의 데이터를 취득한다. (b-d) 회전각이 0인 가우시안 집합 G_0를 회전축 v를 기준으로 회전각 θ_t만큼 회전시킨 가우시안 집합 G_t를 렌더링한 결과와 ground-truth 사이의 손실값을 바탕으로 G_0, v, 그리고 회전각 잔차 예측 모델의 변수 ϵ를 학습한다.

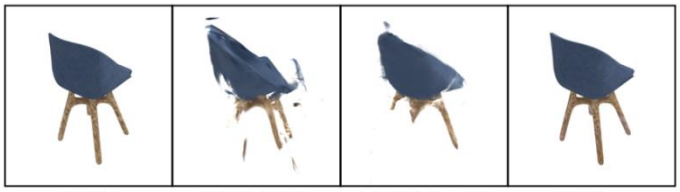


그림 2: 합성 데이터를 활용한 실험 결과, (a) GT image, (b-c) 비교하는 모델의 결과, (d) 제안한 모델의 결과.

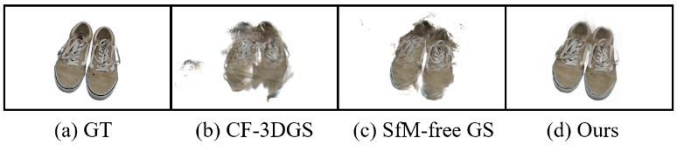


그림 3: 실제 데이터를 활용한 실험 결과.

Model	PSNR↑	SSIM↑	LPIPS↓
CF-3DGS	17.417	0.904	0.137
SfM-free GS	19.584	0.924	0.115
Ours	**37.802**	**0.985**	**0.031**

표 1: 제안한 모델과 비교모델의 정량적 성능 비교 결과.

2.2. 가우시안 속성 초기화와 깊이 감독을 통한 정규화

초기의 부정확한 원판의 회전값에 의해 가우시안 속성이 과적합되는 문제가 발생할 수 있다. 본 논문에서는 이를 해결하기 위해 두 가지 정규화 기법을 도입한다. 첫번째는 깊이 일관성 손실로, 단안 깊이 추정 모델로 얻은 깊이 맵과 3DGS를 렌더링하여 얻은 깊이 맵 간의 차이를 줄여 기하학적 일관성을 강화하고 회전축이 올바른 방향으로 정렬되도록 한다. 두번째는 가우시안 속성 초기화 전략으로, 일정 주기마다 가우시안의 크기와 회전 값을 기본값으로 초기화하여 모델이 부정확한 회전 정보에 과적합되어 가우시안 포인트가 비정상적으로 커지거나 잘못된 방향으로 고정되는 현상을 방지한다.

3. 실험 및 결과

실험을 위해 블렌더를 활용하여 가상 데이터를 생성하였다. 이때 실제 환경을 모사하기 위해 각 이미지의 회전각에 작은 가우시안 노이즈를 추가하여, 회전각 오차를 재현하였다. 표 1에서와 같이, 제안한 모델은 기존 모델 대비 우수한 성능을 보였으며, 이는 기존 모델이 카메라 변수를 정확하게 추정하지 못한 반면, 제안한 모델은 회전 원판에서 촬영했다는 사전 지식을 활용하여 회전축과 회전각을 보다 정확하게 추정하기 때문이다. 이러한 차이는 그림 2의 정성적 실험 결과에서도 확인할 수 있다. 추가적으로 제안한 모델의 각 구성 요소들의 효과를 검증하기 위해 표 2와 같이 ablation study를 수행하였다. 마지막으로, 실제 촬영한 데이터를 사용한 그림 3의 결과를 통해 제안한 방법이 실제 환경에도 효과적으로 적용됨을 보였다.

Ours	w/o 회전축 학습	w/o 회전각 잔차 학습	w/o 깊이 일관성 손실	w/o 가우시안 속성 초기화
27.01	24.22	26.62	25.75	25.81

표 2: 제안한 방법의 효과를 PSNR 지표로 비교한 결과.

4. 결론

본 연구는 고정된 카메라로 회전 원판 위의 물체를 촬영한 이미지로부터 Structure-from-Motion(SfM) 과정 없이 3D 객체를 복원하는 새로운 방법을 제안한다. 다양한 실험을 통해 원판을 활용한 3D 스캔 환경에서 제안된 방법이 기존 방식보다 더욱 정확한 3D 객체 복원이 가능함을 입증하였다. 그러나 본 방법은 카메라의 내부 파라미터를 추정하지 않는다는 한계가 있다. 향후, 카메라 내부 파라미터 추정을 포함한 확장 연구를 통해 보다 다양한 분야에 활용될 수 있을 것으로 기대한다.

참고문헌

[1] Kerbl, et al. 3d gaussian splatting for real-time radiance field rendering. In ACM Trans. Graph, 2023
[2] Schonberger, et al. Structure-from-motion revisited. In CVPR, 2016
[3] Fu, et al. Colmap-free 3d gaussian splatting. In CVPR, 2024
[4] Ji, et al. SfM-Free 3D Gaussian Splatting via Hierarchical Training. In CVPR, 2025

CEM 을 활용한 확산 기반 초해상도 모델 성능 개선*

정상준, 김재환, 최준서, 정승화 [1]
몰입형 미디어 연구실
{2002sangjun@naver.com, jonny0010@naver.com, wnstj010126@gmail.com, seunghwajeong@sejong.ac.kr}

Performance Enhancement of Diffusion-based Super-Resolution Models Using CEM

Sang Jun Jeong, Jae hwan Kim, Jun seo Choi, Seong hwa Jeong
Immersive Media Laboratory

요약

확산 (Diffusion) 기반의 초해상화 (Super Resolution, SR) 모델은 높은 품질의 고해상도 이미지를 제공하지만, 원본 구조의 변형 문제가 발생한다. 본 논문에서는 일관성 유지 모듈(CEM)을 사용하여 확산 기반의 SR 모델의 품질을 유지하면서 원본 구조의 변형을 최소화하는 방법을 제안한다. 확산 기반 SR 의 대표 모델인 StableSR 모델에서 CEM 모듈의 비교 실험을 통해, 초해상화 모델의 품질을 유지하면서 원본의 정확도 향상할 수 있음을 정량적으로 입증하였다.

1. 서론

최근 디퓨전(Diffusion)은 우수한 성능을 기반으로 다양한 분야에서 활발히 활용되고 있다. 특히, 이미지 초해상도(Super-Resolution, SR)는 디퓨전 모델이 효과적으로 적용되고 있는 대표적인 응용 분야 중 하나이다. 그러나 디퓨전 기반 단일 이미지 초해상화(Single Image Super-Resolution, SISR)는 시각적으로 자연스러운 이미지를 생성하는데 초점이 맞춘 모델이기 때문에 원본 이미지의 구조적 특징을 정확하게 반영하지 못하는 한계를 지닌다. 이에 본 논문에서는 이러한 문제를 해결하기 위해, 원본 이미지와 출력 이미지의 구조적 일관성을 유지시키는 일관성 유지 모듈(Consistency-Enforcing Module, CEM)을 도입하여 원본 저해상도 이미지의 구조적 정보를 최대한 보존하면서도 높은 품질의 초해상도 이미지를 생성하는 방법을 제안한다.

2. 배경 연구

2.1. StableSR

본 연구는 확산 기반의 단일 이미지 SR 의 대표 모델인 StableSR[1]을 사용하였다. 기존 Stable Diffusion 모델의 파라미터는 그대로 유지하고, 소수 학습 가능한 파라미터만 파인 튜닝하여 생성 능력을 초해상도에 최적화한 모델이다.

확산 기반 모델은 확률적 특성 때문에, 디코더에서 추출한 정보(Fd)만을 이용하면 시각적 품질은 높지만, 원본 이미지의 구조를 유지할 수 없다. 반대로 저해상도 정보를 지나치게 활용하면 품질 향상이 어려운 Trade-off 가 발생한다. StableSR 은 CFW(Controllable Feature Wrapping) 모듈을 통해 가중치 □값을 통해 원본 이미지를 반영하는 비율을 사용자가 원하는 목적에 부합하게 생성하도록 제공한다. 본 연구에서는 확산 모델의 장점을 유지하면서 CEM[2]을 적용하여 원본 해상도의 구조를 보존하여, StableSR 이 가지는 Trade-off 를 최소화한다.

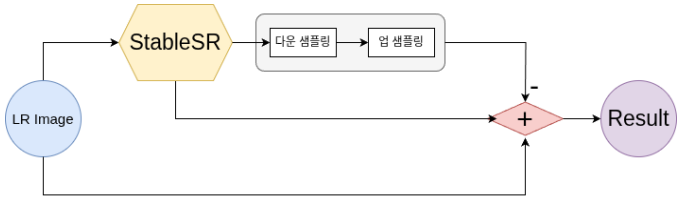


그림 1. CEM 모듈의 구조 [2]

2.2. 일관성 유지 모듈

CEM[2]은 저주파수 영역은 원본 영역을 유지하면서 고주파수 영역은 SR 결과물의 품질을 유지하는 모듈이다. 그림 1 과 같이 SR 로 얻은 결과물을 다운 및 업 샘플링을 수행한 결과값과 차이 값을 계산하여 고주파수 영역을 추출한다. 마지막으로 원본 저해상도 이미지를 업 샘플링한 결과와 합산하여 원본 구조를 유지하는 방법이다.

본 연구에서는 제안된 CEM 의 연구결과를 구현하여 StableSR 의 결과물에 적용하였다.

3. 실험 방법

본 연구는 기존 StableSR 을 통해 생성된 초해상도(SR)

---

1 교신저자
* 포스터발표논문
* 학부생 주저자 논문임

이미지와, 일관성 유지 모듈 CEM)을 적용하여 구조적 특징을 보강한 SR 이미지를 비교하였다. CFW 모듈의 가중치 w를 통해 두 정보 중 어느 쪽에 더 집중할지를 설정한다. w=0 일 경우 디코더 feature 에만 집중하며, w=1 일 경우 저해상도 입력에만 의존한다.

w	PSNR	NIQE	MUSIQ	CLIPIQA	BRISQUE
0.0	19.60	3.39	73.89	0.72	11.16
0.1	19.52	3.31	73.92	0.71	**10.41**
0.2	**19.70**	**3.24**	73.92	**0.74**	11.76
0.3	19.23	3.36	**73.96**	0.73	11.84
0.4	19.01	3.45	73.87	0.72	12.66
0.5	18.76	3.62	73.70	0.71	13.40

[표 1] w값에 따른 이미지 수치 비교

비교군인 StableSR[1] 결과물은 논문에서 성능이 가장 우수 설정으로 보고된 w=0.5 를 사용하였다. 반면, CEM 을 적용한 본 연구의 결과물은 구조적 특징을 강화함에 따라 더 낮은 값을 사용하였고, 실험을 통해 높은 정확도(PSNR)을 유지하면서 인지적 품질이 가장 좋은 w값을 사용하였다. NIQE 및 BRISQUE 는 낮을수록, MUSIQ 와 CLIPIQA 는 높을수록 높은 인지적 품질임을 수치로 나타낸 값이다. [표 1]에 표시된 실험 결과를 바탕으로 본 실험에서는 w=0.2 을 선택하여 비교 연구를 수행하였다. 테스트는 DIV2K 학습 데이터셋의 10 개의 이미지를 랜덤하게 선택하였으며, 적용 결과를 정량적으로 비교하였다.

그림 2. 좌측부터 LR 이미지, CEM 미적용 이미지, 적용 이미지

CEM 적용여부	PSNR	NIQE	MUSIQ	CLIPIQA	BRISQUE
미적용	18.76	3.73	73.70	**0.76**	12.57
적용	**19.70**	**3.24**	**73.92**	0.74	**11.76**

[표 2]각 이미지 별 인지적 품질 비교

3.1. 평가

두 이미지의 PSNR 비교를 통해 CEM 이 원본 이미지의 구조적 특징을 보강하였는지 확인했다. PSNR 을 비교하면 CEM 이 적용된 이미지가 PSNR 지표에서 더 높은 값을 보이는 것으로 나타났다. 이는 CEM 이 원본 이미지의 구조적 특징을 효과적으로 반영하여, 실제 이미지 복원 성능 향상에 기여함을 의미한다. 그림 2 의 바구니 영역과 같이, CEM 을 미적용한 이미지는 저해상도 이미지에서 확인 가능한 십자 구조가 사라졌지만 CEM 을 적용한 이미지에서는 보다 선명하게 구조가 유지되고 있음을 시각적으로 확인했다. CEM 모듈 내, SR 과정에서 왜곡된 구조적 특징을 제거하고 고주파수 영역만 유지하고, 저주파수 영역은 원본을 유지함으로써, 이미지의 변형을 최소화 하였다.

[표 2]와 같이, 인지적인 이미지의 품질을 측정하는 지표들 (NIQE, MUSIQ, CLIPIQA 및 BRISQUE)을 활용하여 CEM 미적용 및 적용 결과를 비교 분석하였다. 일부 항목에서는 CEM 미적용 모델이 우위를 보였으나, 다른 항목에서는 CEM 적용 모델이 더 나은 성능을 보였으며, 모든 지표에서 품질의 차이가 미미하다. 즉, 기존 확산 기반 모델 결과물의 품질을 해치지 않음을 확인하였다.

종합적으로 확산기반 모델을 통해 생성된 높은 품질의 SR 결과물을 유지하였으며, 원본 이미지의 구조적 특징을 생성된 이미지에 적용하여 정확도가 개선됨을 정량적 지표를 통해 확인하였다.

4. 결론

본 연구는 확산 기반 SR 모델의 출력 이미지를 CEM 을 활용하여, 시각적 품질을 저해하지 않으면서 복원 정확도를 향상시키는 방법을 제안하였다. 확산 기반 SR 의 대표 모델인 StableSR 모델에서 CEM 모듈의 비교 실험을 통해, 초해상화 모델의 품질을 유지하면서 원본의 정확도를 향상시킬 수 있었다.

참고문헌

[1]Wang, J., Yue, Z., Zhou, S., Chan, K. C. K., & Loy, C. C. (2024). Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015.
[2]Bahat, Y., & Michaeli, T. (2020). Explorable super resolution. arXiv preprint arXiv:1912.01839.
[3]Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.
[4]Han, Y., Yu, T., Yu, X., Wang, Y., & Dai, Q. (2023). Super-NeRF: View-consistent detail generation for NeRF super-resolution. arXiv

preprint arXiv:2304.13518.

2D 이미지 분할을 이용한 텍스트 기반 3D 부분 텍스처 편집*

이다예[0], 이상은[0], 정지우[0], 김영준[1]
이화여자대학교 컴퓨터공학과
{tit101010, sang_rlo, jjw21, kimy}@ewha.ac.kr

Text-guided Part-Level Texture Editing using 2D Image Segmentation

Daye Lee[0], Sangeun Lee[0], Jiwoo Jung[0], Young J. Kim[1]
Dept. of Computer Science and Engineering, Ewha Womans University

그림 1: 자연어 기반 텍스처 편집 결과. 좌측부터 입력 이미지, 자연어 명령어, 텍스처가 적용된 메쉬 렌더링.

요약

본 연구는 자연어 프롬프트를 기반으로 3D 객체의 특정 부위에 원하는 텍스처를 적용할 수 있는 경량화된 파이프라인을 제안한다. 제안하는 방법은 2D 이미지 분할과 마스크 기반 조건부 인페인팅을 활용한다. 기존 3D 기반 메시 분할 및 텍스처 편집 기법에 비해 연산 비용이 낮고, 별도의 학습 없이 효율적으로 부위별 편집이 가능하다. 비교 실험을 통해 부위별 텍스처 변경이 가능하며, 기존 방식 대비 표현 정확도가 향상되었음을 확인하였다.

1. 서론

자연어를 활용한 세밀한 스타일 편집은 2D 이미지에서 사용자 접근성을 크게 향상시켰으며, 이러한 기능에 대한 수요는 3D 환경에서도 증가하고 있다. 기존 연구[1][2][3]들은 주로 객체 전체의 텍스처 편집에 초점을 맞추고 있어, 세부 부위 편집을 위해서는 별도의 3D 메쉬 분할 기법이 필요하다. 그러나 3D 메쉬 분할[4]은 학습 및 연산 비용이 높기 때문에 실제 응용에 제약이 있다. 본 연구에서는 자연어 프롬프트를 기반으로 2D 이미지 공간에서 부위별 텍스처를 편집한 후, 이를 3D 메쉬의 특정 부위에 통합하는 경량화된 텍스처 편집 파이프라인을 제안한다.

* 포스터 발표, 학부생 주저자 논문
[0] 공동 주저자
[1] ITRC/IITP 프로그램 IITP-2023-2020-0-01460와 연구재단 2022R1A2B5B03001385의 지원

2. 3D 부분 텍스처링 파이프라인

그림 2는 제안하는 텍스처 편집 파이프라인의 구조를 나타낸다. 입력 이미지는 사용자가 3D로 생성하고자 하는 이미지이거나, 편집 대상 3D 모델의 특정 시점 캡처 이미지이다. 자연어 텍스트로 지정된 부위 정보를 바탕으로, 2D 이미지 상에서 Grounded-SAM[5]을 활용해 해당 부위를 분할하고 부위별 마스크를 생성한다. 이후, ControlNet[6] 기반 조건부 인페인팅을 통해 지정 부위에 텍스처링을 적용한다. 텍스처링된 이미지는 Instant-Mesh[7]를 통해 멀티뷰 생성 후 최종 3D 메쉬로 재구성된다. 이를 통해 별도의 3D 메쉬 분할 없이도 세부 부위에 대한 텍스처 편집이 가능하다.

2.1 사용자 명령 프롬프트 최적화 및 이미지 전처리

사용자의 자연어 프롬프트는 LLM을 활용해 처리한다. 가장 먼저 명령어는 Object-Part-Style 구조로 분할된다. 이를 기반으로 다음 단계에서 스타일 프롬프트의 묘사를 구체화하여, inpainting 과정에서 정밀한 제어가 가능하도록 한다. 입력 이미지는 부위 인식 및 inpainting 성능 향상을 위해 배경 제거와 업스케일링 전처리를 수행한다.

2.2 텍스트 기반 2D 편집 영역 인식

전처리된 입력 이미지로부터 부위 마스크를 생성하기 위해 Grounded-SAM[5]을 사용한다. 텍스트 프롬프트는 "부위, 객체" 형식으로 반복 구성하여 토큰화 오류를 줄이고, 객체 맥락을 반영하도록 한다. 또한, 부위 대신 전체 객체가 더 높은 점수로 탐지되는 오류를 줄이기

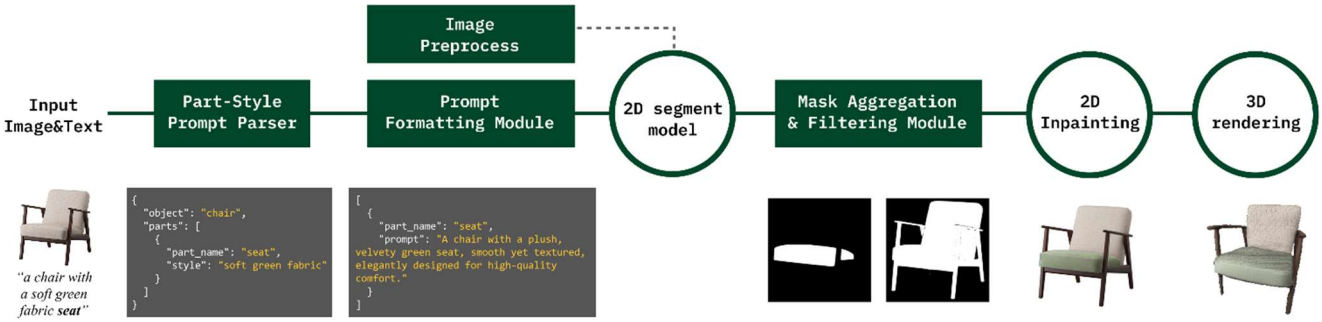


그림 2: 전체 파이프라인과 입력 처리 예시.

위해, 후보 마스크들을 객체 단위 마스크와 유사도 기반으로 비교해 필터링하는 후처리 로직을 적용한다.

2.3 텍스트 기반 인페인팅

텍스처 변경은 Stable Diffusion[8] 기반의 inpainting 모델과 ControlNet[6]을 결합하여 수행한다. 이때, 부위 및 스타일 정보를 반영하여 재가공한 프롬프트를 활용한다. 또한, 불필요한 객체 생성이나 배경 변형을 억제하기 위해 부정 프롬프트를 적용하며, 하이퍼 파라미터를 실험적으로 조정하여 구조 보존성과 표현력 간의 균형을 최적화한다.

2.4 3D 렌더링 및 텍스처 생성

기존의 NeRF 기반 3D 렌더링 기법은 높은 연산 비용과 학습 과정으로 인해, 실시간 처리 및 경량화 요구를 충족하기 어렵다. 이에 본 연구에서는 Instant-Mesh[7]를 활용하여 편집된 결과를 3D 메쉬로 재구성하고 2D 상의 변형이 적용된 텍스처 맵을 생성한다.

3. 실험

3.1 환경 및 설계

실험은 NVIDIA RTX 4090, Ubuntu 20.04 환경에서 수행하였다. 실험 대상 객체 범주는 총 5종(의자, 조명, 컵, 테이블, 인형)으로 구성하였다. 실험에 사용된 3D 메쉬는 공개된 모델 중 부위 편집이 가능한 객체를 우선 선정하였으며, 추가로 Blender를 활용해 부위별 텍스처 변형이 된 형태로 제작하였다. 이러한 모델을 기반으로 텍스처 변경 프롬프트를 설계하여 총 15개의 실험 조건을 구성하였다.

텍스처링 비교 대상으로는 Paint3D[1], TEXTure[2], SyncMVD[3]를 선정하였다. 생성된 메쉬 결과를 360도 회전시켜 다각도 영상을 생성한 뒤, 최대 30프레임(해상도 512×512)을 추출하여 비교하였다.

정량 평가는 색상 차이(ΔE, CIEDE2000), 시각적 유사도(LPIPS, AlexNet 기반), 텍스처 대비 차이(GLCM Contrast Difference), 색상 분포 유사도(Histogram Similarity, RGB 채널별 1D 히스토그램 기반 상관계수)의 네 가지 지표로 수행하였다. 각 지표는 프레임별로 계산한 후 평균값으로 비교하였다.

3.2 결과

본 연구에서 제안한 파이프라인은 모든 정량 지표에서 baseline 대비 가장 우수한 성능을 보였다. 색상 유사도(ΔE)는 2.75로 가장 낮았고, LPIPS(0.112) 및 GLCM Contrast Difference(22.87) 또한 가장 낮아, ground truth와의 통계적·인지적 유사도가 가장 높았다.

표 1: 정량적 실험 결과.

모델	ΔE	LPIPS	GLCM	Histogram
Ours	2.75	0.11	22.87	0.9999
Paint3D	3.79	0.14	33.19	0.9998
TEXTure	4.04	0.16	39.23	0.9998
SyncMVD	3.72	0.14	36.26	0.9998

그림 3은 프롬프트 "a chair with a soft green fabric seat"에 대한 결과를 시각적으로 비교한 예시이다. Baseline 모델들은 세부 부위별 지시를 정확히 반영하지 못한 반면, 제안한 파이프라인은 사용자가 지정한 부위에 정확하게 텍스처를 적용하였다.

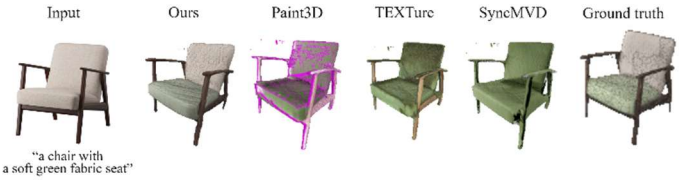


그림 3: 실험 결과의 시각적 비교.

4. 결론

본 연구에서는 자연어 입력을 기반으로 보다 직관적인 3D 텍스처 편집 파이프라인을 제안하였다. 2D 기반 부위 인식 및 텍스처링 기법을 효과적으로 적용하고 최적화하여 경량화된 파이프라인을 구현하였다.

향후 본 기법을 후처리 모듈 및 3D 디자인 도구와 통합할 경우, 더 높은 품질의 결과를 생성할 수 있으며, 일반 사용자의 3D 콘텐츠 제작 참여를 촉진하는 데에도 기여할 수 있을 것이다.

참고문헌

[1] Zeng X., Chen X., Qi Z., Liu W., Zhao Z., Wang Z., Fu B., Liu Y., Yu G., Paint3D: Paint Anything 3D with Lighting-Less Texture Diffusion Models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4252–4262, 2024.

[2] Richardson E., Metzer G., Alaluf Y., Giryes R., Cohen-Or D., TEXTure: Text-Guided Texturing of 3D Shapes, *ACM*

Transactions on Graphics (TOG), 42(6):1–10, 2023.
[3] Liu Y., Xie M., Liu H., Wong T., Text-Guided Texturing by Synchronized Multi-View Diffusion, *ACM SIGGRAPH Asia 2024 Conference Proceedings (SA '24)*, Article 60, pp. 60:1–60:11, 2024.
[4] A. Abdelreheem, I. Skorokhodov, M. Ovsjanikov, and P. Wonka, SATR: Zero-Shot Semantic Segmentation of 3D Shapes, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15120–15133, 2023.
[5] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, *arXiv preprint arXiv:2401.14159*, 2024.
[6] L. Zhang, A. Rao, and M. Agrawala, Adding Conditional Control to Text-to-Image Diffusion Models, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023.
[7] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models, *arXiv preprint arXiv:2404.07191*, 2024.
[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

혼합현실 소방훈련 시뮬레이터: HMD와 IoT 소화기를 활용한 체화 학습

이순교[0], 김필중, 전종민, 김예은, 최수미*
세종대학교 컴퓨터공학과, 초실감XR연구센터
{sungyo21c, fjfo101, klm6564}@sju.ac.kr, {kyy1462, smchoi}@sejong.ac.kr

Mixed Reality Firefighting Simulator: Embodied Learning using HMD and IoT Fire Extinguisher

Sungyo Lee[0], Phil-Joong Kim, Jong-Min Jeon, Ye-Eun Kim, Soo-Mi Choi*
Department of Computer Science and Engineering and the XR Research Center, Sejong University

요약

본 논문은 이론 중심 소방안전 교육의 몰입도와 현실감 부족 문제를 해결하기 위해, IoT 센서를 탑재한 소화기 모형과 혼합현실(Mixed Reality) 기술을 활용한 훈련 시뮬레이터를 개발하였다. 사용자는 혼합현실 환경에서 실제 소화기를 조작하며 화재를 진압하는 능동적인 훈련을 통해, 단순한 인지 학습을 넘어 신체 기억을 유도하는 체화 학습(Embodied Learning)이 가능하다.

1. 서론

최근 연이어 발생한 대형 화재사고로 인해 소방안전 교육의 중요성이 다시금 주목받고 있다. 기존의 이론 중심이거나 수동적인 소방안전 교육은 화재 상황에서 요구되는 구체적 행동 요령을 체득하기 어려우며, 가상현실(Virtual Reality) 기반 교육조차도 장비 조작이나 촉각 피드백 등 물리적 상호작용이 결여되어 현실감 측면에서 한계를 지닌다. 이에 따라 실제 장비 조작을 포함한 실습형 교육의 필요성이 제기되고 있으며, 이러한 맥락에서 디지털트윈과 실감형 기술을 활용한 혼합현실 기반 접근 방식이 새로운 대안으로 떠오르고 있다[1].
본 연구는 IoT 장치를 탑재한 소화기 모형과 ZED 깊이 카메라를 활용해 현실 공간을 인식하고 화재 시나리오를 구현하는 혼합현실 소방훈련 시스템을 제안한다. 이를 통해 사용자는 몰입감 있는 실습 환경에서 소화기를 직접 조작하며 화재 진압 경험을 쌓아 효과적인 대응 능력을 익힐 수 있다.

2. 시스템 설계 및 구현

2.1. IoT 기능을 탑재한 소화기

제안 시스템은 교육용 물 소화기(4L)에 Arduino UNO 보드를 탑재하여, 사용자의 물리적 조작 정보를 혼합현실 환경에서 디지털 신호로 변환·처리하도록 설계하였다.
그림 1의 (a)는 본 시스템에서 가상 및 현실 요소를 담당하는 기술 구성과 이를 동기화하는 전체 파이프라인이다. (b)는 IoT 구현을 위한 하드웨어의 구성으로 버튼 모듈, 오디오 잭 모듈, MPU-9250 관성 측정 장치(Inertial Measurement Unit, IMU)가 포함된다. 버튼 모듈은 소화기 손잡이 내부에 장착되어 있어, 사용자가 레버를 당길 때 버튼 스위치가 눌리는 방식이다. 오디오 잭 모듈은 안전핀의 삽입 여부를 감지한다. 안전핀이 제거되지 않은 상태에서는 소화기 작동이 차단되도록 설계하였다. 이는 실제 화재 상황에서 안전핀 제거를 간과하는 문제를 방지하기 위한 것으로, 안전핀 제거 단계의 중요성을 강조하는 교육적 목적을 반영한 설계이다. MPU-9250은 자이로스코프, 가속도계, 지자기 센서가 통합된 9축 IMU 센서로, 사용자의 손 움직임에 따른 roll, pitch, yaw 값과 x, y, z 방향의 각속도를 계산하여, 소화기 호스의 방향성과 위치를 정밀하게 추적하는 데 활용된다. [2]. 이러한 입력 정보는 Arduino IDE를 통해 실시간으로 처리되며, 시리얼 통신을 통해 Unity 환경으로 전달된다.
Unity 엔진에서는 C# 스크립트를 활용하여 각 센서로부터 수신된 신호를 가상 시나리오로 동적 생성하며, 사용자의 실제 조작 동작이 혼합현실 내 화재 진압 행동으로 실시간 반영되도록 구성하였다. 현실의 물리적 상호작용과 가상 환경 간의 연동성을 강화하여, 보다 몰입도 높은 혼합현실 기반 소방훈련 경험을 제공할 수 있도록 하였다. (c)는 혼합현실 환경에서 가상의 불과 현실의 소화기 간 통합이 어떻게 이루어졌는지 시각적으로 보여준다.

2.2. HMD와 깊이 카메라로 혼합현실 환경 구축

Meta Quest 3는 내장 카메라 특성상 Unity 내에서 개발자가 직접 활용할 수 있는 정밀한 깊이 정보를 제공하지 않는다. 이는 가상 콘텐츠와 현실 공간을 정합해야

---

* 포스터 발표논문
* Corresponding author: smchoi@sejong.ac.kr
* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT연구센터지원사업(IITP-2025-RS-2022-00156354)의 지원을 받아 수행된 연구임.

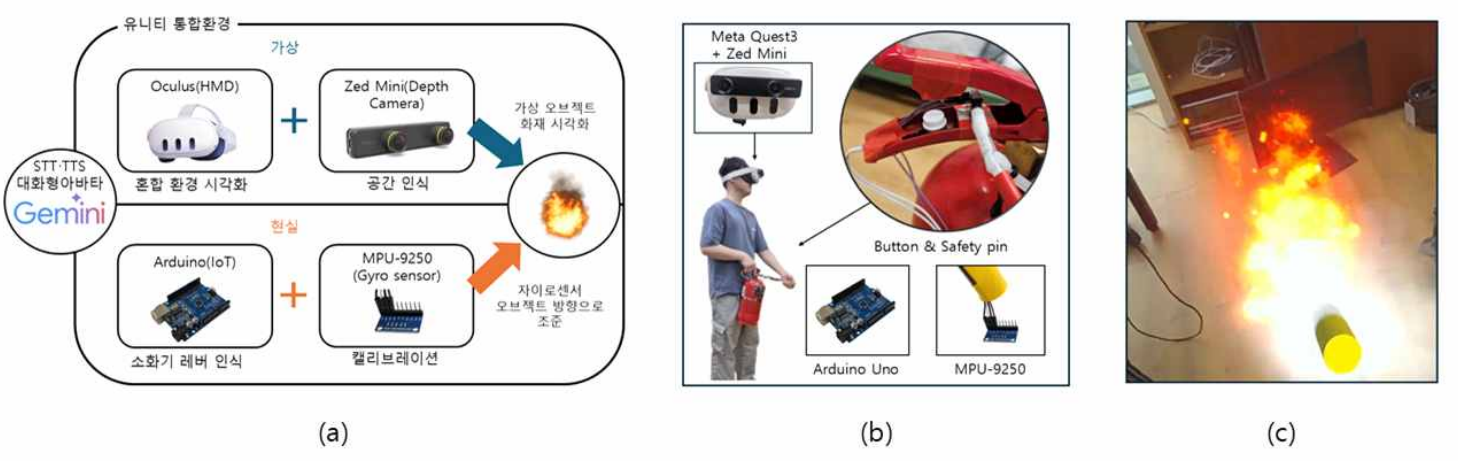


그림 1: 실가상 통합 소방 훈련 시스템, (a) 시스템 아키텍처, (b) 주요 하드웨어 구성, (c) HMD 시뮬레이션 화면

하는 혼합현실 콘텐츠 개발에 있어 제약으로 작용한다. 이에 본 연구에서는 HMD 전면에 깊이 카메라인 ZED Mini를 부착하여 스테레오 비전 기반의 깊이 센싱 기술을 통해 정밀한 실시간 공간 인식이 가능하다. ZED Mini는 사용자의 위치, 벽체, 바닥 평면 등을 인식하여, 실제 공간 조건에 맞는 위치에 가상 화재 상황을 생성하고, 사용자의 시선 및 동작에 따라 상호작용이 이루어지도록 설계하였다.

이를 통해 사용자는 단순한 가상 체험을 넘어서, 현실 공간에 기반한 실감형 훈련 시나리오를 경험할 수 있다. 화재 발생 위치와 대응 동작 간의 공간적 일치로 몰입도와 학습 효과를 동시에 강화할 수 있다. 시스템 구현에는 Meta XR SDK 중 Meta Interaction SDK와 Passthrough API를 적용하여 Unity 기반의 혼합현실 콘텐츠로 개발하였으며, 실제 소화기 조작 또한 실시간으로 반영될 수 있도록 구성하였다.

2.3 혼합현실 환경에서의 화재진압 콘텐츠 개발

콘텐츠 체험에 앞서 참여자에게 실제 소화기 조작법을 가상의 소방 아바타가 안내한다. 아바타는 음성 합성 기술(TTS)을 통해 소화기의 손잡이와 호스를 올바르게 잡는 방법, 안전핀 제거에 대한 설명을 제공하며, 이 과정에서 물리 장비에 장착된 센서의 작동 방식과 역할에 대해서도 간략히 안내한다. 안전핀이 제거되어야만 화재 진압 동작을 인식하도록 설계하였기 때문에, 조작 절차의 정확한 습득은 체험의 필수 전제가 된다.

이어서 참여자는 Meta Quest 3 HMD를 착용한 상태에서 혼합현실 환경에 진입하게 된다. 시스템은 ZED Mini를 통해 화재 상황이 적절한 위치에 생성된다. 참여자는 주변을 탐색하면서 불꽃을 직접 발견하고, 실제 소화기를 조작하여 화재 진화 동작을 수행하게 된다.

그림 2: 혼합현실 환경에서 화재를 진압하는 어린이 모습

그림 2는 혼합현실 환경에서 IoT 기능이 탑재된 교육용 물 소화기로 화재를 진압하는 어린이의 모습을 보여준다. 이 소화기는 약 4kg의 무게로 어린이에게는 다소 무거워, 자연스럽게 다리로 받쳐 들게 되는 모습을 확인할 수 있다.

3. 결론 및 향후 연구

본 연구는 기존 가상 소화기 교육훈련에서 나타나는 참여도와 몰입도의 한계를 극복하고자, 실물 기반 IoT 장치와 가상현실 기술을 융합한 혼합현실 기반 훈련 시스템을 설계 및 구현하였다.

완벽한 3차원 위치 추적을 구현하려면 자이로 센서와 함께 개별 외부 트래커가 필요하기에 추가하여 보완할 계획이다.

향후 연구에서는 소방 관련 질의응답 및 탈출 시나리오 진행을 위해 대규모 언어모델(LLM) 기반의 음성 인식(STT) 및 음성 합성 기술을 연동된 대화형 아바타 인터페이스로 발전시킬 계획이다[3]. 이를 통해, 사용자는 자연어 기반 음성 상호작용만으로도 가상환경과 현실 장치 간의 유기적 통합을 경험할 수 있으며, 이는 향후 몰입형 교육훈련 시스템의 핵심 요소로 작용할 수 있을 것으로 기대된다.

참고문헌

[1] Park Chan Gon, Effects of Hands-on Safety Education Program on High-School Students' Fire-Safety Awareness, National Fire Research Institute of Korea, Vol.5:54-70, 2024.

[2] Johnny Leporcq, Position Estimation Using an Inertial Measurement Unit Without Tracking System, Master's Thesis, Aalto University, School of Electrical Engineering, Espoo, 2018.

[3] Gemini Team Google, Gemini: A Family of Highly Capable Multimodal Models, arXiv 2025.

사용자 참여형 얼굴 스타일 이미지 데이터셋 생성 방법*

이영균[0], 김장호, 김준호
국민대학교 컴퓨터공학과
{yglee981130, jangho.kim, junho}@kookmin.ac.kr

Constructing Stylized Face Image Datasets via Human Evaluation

Younggyun Lee[0], Jangho Kim, Junho Kim
Dept of Computer Science, Kookmin University

요약

본 연구는 사용자 평가 기반의 정량적 기준을 활용하여 원본 인물의 정체성과 스타일 표현이 균형 있게 반영된 얼굴 스타일 이미지 데이터셋을 구축하는 방법론을 제안한다. CelebA-HQ 이미지에 대해 다양한 스타일(픽사, 지브리)을 적용하여 생성한 후보 이미지들을 바탕으로 사용자 연구를 실시하고, 원본 인물의 정체성을 유지하면서 스타일 표현이 우수한 이미지를 선별하였다. 이후, 선택된 이미지들의 LPIPS (Learned Perceptual Image Patch Similarity) [1] 값을 분석해 정체성과 스타일 간 균형을 수치적으로 정의하고, 이를 통해 최적의 이미지 생성 조건을 역추정하였다. 추가 사용자 평가를 통해 해당 기준의 타당성을 검증함으로써, 제안한 정량적 기준이 스타일 이미지 데이터셋 구축에 효과적으로 활용될 수 있음을 확인하였다.

1. 서론

최근 인공지능 기술의 발전으로, 실제 인물의 얼굴 사진을 만화나 그림처럼 스타일화하는 기술이 대중적으로 활용되고 있다. 그러나 이러한 기술은 타인의 얼굴을 무단으로 스타일화해 악용되는 사례에 대한 우려도 함께 낳고 있다. 이러한 폐해를 방지하려면, 스타일 이미지로부터 원본 인물의 정체성을 판별할 수 있는 인공지능 기술이 필요하며, 이에 앞서 훈련 데이터셋으로서 정체성 보존 여부가 명확히 정의된 대량의 (원본, 스타일) 이미지 쌍의 수집이 선행되어야 한다.

본 논문에서는 정체성과 스타일 표현이 균형 있게 반영된 이미지 쌍으로 구성된 대규모 데이터셋을 구축하는 방법론을 제안한다. 소수의 원본 이미지에 대해 다양한 하이퍼파라미터 설정으로 생성된 다수의 스타일 이미지들을 대상으로 사용자 연구를 수행하고, 이를 통해 정체성과 스타일 표현을 모두 만족하는 이미지의 특성을 분석하였다. 이 과정에서 LPIPS (Learned Perceptual Image Patch Similarity) [1] 기반의 정량적 기준을 도출하고, 이를 바탕으로 이미지 생성 모델의 최적 하이퍼파라미터를 역추정하였다. 이후 역추정된 하이퍼파라미터로 대량 생성한 스타일 이미지가 정체성 보존과 스타일 표현이라는 두 가지 기준을 모두 만족하는지를 추가 사용자 평가를 통해 검증함으로써, 제안한 방법의 실효성을 입증하였다.

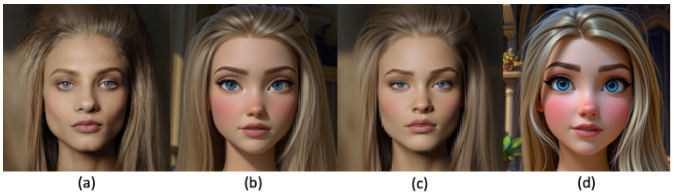


그림 1. 원본 이미지와 스타일 이미지
(a) 원본, (b) 정체성과 스타일 표현이 균형 잡힌 이미지, (c) 스타일이 부족한 이미지, (d) 정체성이 손실된 이미지

2. 실험

본 연구에서는 그림 1.(b) 와 같이 원본 얼굴 이미지에 대해 “원본 인물의 정체성을 유지하면서도 주어진 스타일이 최대한 적용된 상태”의 합성 이미지를 스타일 이미지의 기준으로 정의한다. 이러한 기준을 정량적으로 측정하기 위해 다음과 같은 실험 설계를 구성하였다.

백본 생성 모델 CelebA-HQ 데이터셋을 기반으로 얼굴 스타일 이미지 데이터를 생성하였다. 스타일 생성에는 FLUX 기반의 Image-to-Image 변환 모델을 사용하였으며, 실험은 3차원 픽사(Pixar) 애니메이션 캐릭터 스타일과 지브리(Ghibli) 스튜디오 애니메이션 캐릭터 스타일을 주요 대상으로 설정하였다. 스타일 모듈은 LoRA를 통해 사전 학습된 파라미터를 적용하여 모델에 주입하였다.

이미지 생성 과정에서 원본 인물의 구조적 정보를 보존하기 위해 두 가지 보조 모듈들을 활용하였다. ControlNet은 얼굴의 기하 구조를 명시적으로 모델에 제공하여 인물 형태의 일관성을 유지하도록 하였으며, Florence2는 원본 이미지로부터 의미 기반 텍스트 프롬프트를 생성하여 생성 모델의 조건 입력으로 활용하였다. 이를 통해 이미지별 세밀하고 자연스러운 결과를 생성할 수 있도록 하였다.

* 포스터 발표논문
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1F1A1074628).

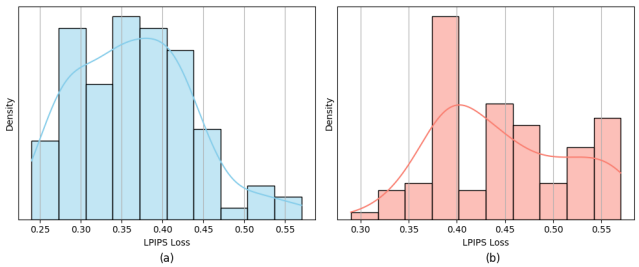


그림 2. LPIPS 손실 분포 (a)는 픽사 스타일, (b)는 지브리 스타일에 대한 LPIPS 손실의 분포.

2.1. 사용자 연구 1단계: 정량적 기준 수립

실험 데이터 구축 CelebA-HQ 데이터셋에서 50장의 원본 이미지를 선정하고, 백본 생성 모델의 주요 하이퍼파라미터를 다음과 같이 조절하여 실험에 활용될 후보 이미지를 생성한다. 백본 생성 모델은 저해상도 잠재 이미지를 생성한 뒤 이를 업스케일링하여 고해상도 이미지를 생성하며, 이때 단계별로 쓰이는 잡음 제거 강도(denoising strength)가 생성 이미지의 스타일화 정도를 결정하는 핵심 변수로 작용한다. 본 실험에서는 0.40부터 0.85까지 0.05 간격으로 값을 변화시키며 스타일 이미지를 생성하는 방식을 사용해, 원본 이미지 1장당 총 100장의 후보 이미지를 생성하였다.

정량적 기준 수립 사용자 연구에는 20~40대의 남성 5명과 여성 5명, 총 10명이 참여하였다. 참가자는 학부생, 대학원생, 일반인 등으로 구성되었고, 모두 스타일 전이에 대한 기술적 사전 지식은 없었다. 참가자들에게는 10명의 원본 인물 이미지와 각각에 대해 스타일이 적용된 100장의 후보 이미지가 제공되었으며, 스타일은 픽사 및 지브리 두 가지 유형으로 나뉘어 제시되었다. 참가자들은 각 원본 이미지에 대해 "정체성이 유지되며 스타일 표현이 가장 우수한" 이미지를 1장씩 선택하였다.

참가자들이 선택한 이미지에 대해 LPIPS 값을 계산하였고, 이를 통해 각 스타일 유형별로 사용자 기준을 만족하는 이미지들의 LPIPS 값에 대한 분포를 수집하였다. 그림 2. 은 두 스타일의 LPIPS 손실 값에 대한 분포 그래프이다. 두 스타일 모두 평균과 분산을 갖는 정규 분포에 근사하는 경향을 보였으며, 평균값이 픽사 스타일은 약 0.366, 지브리 스타일은 약 0.447이고 분산은 두 스타일 모두 0.005인 것으로 확인되었다. 이 분포의 평균값을 해당 스타일의 정량적 기준 대표값으로 정의하였다.

2.2. 사용자 연구 2단계: 정량적 기준 검증

실험 데이터 구축 도출된 LPIPS 기준 분포를 바탕으로, 각 스타일 이미지 생성 모델에 대해 다양한 하이퍼파라미터 설정으로 생성된 이미지 중 LPIPS 값이 대표값에 가장 근접한 경우가 가장 많았던 설정을 최적 하이퍼파라미터로 역추정하였다. 이 하이퍼파라미터는 이후 대규모 데이터셋 구축 시 사용되는 표준 조건으로 채택되었다.

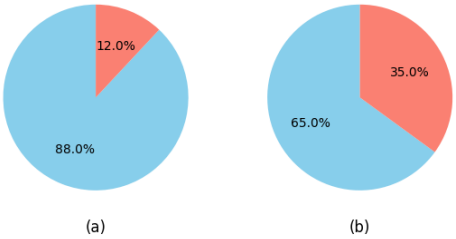


그림 3. LPIPS 기반 정량적 기준 검증 결과 (a)는 픽사 스타일, (b)는 지브리 스타일에 대해 평균 손실 선택 비율. 파란색이 LPIPS 손실의 대표값에 근접한 세트를 선택한 비율.

정량적 기준 검증 역추정된 하이퍼파라미터의 타당성을 검증하기 위해 추가 사용자 연구를 실시하였다. 참가자들은 이전 실험에 사용되지 않은 새로운 인물 이미지와 그에 대해 생성된 스타일 이미지 3장을 제시받았다. 각 세트는 다음 세 가지 LPIPS 손실 수준을 반영하였다: i) LPIPS 손실이 대표값에 근접한 세트, ii) LPIPS 손실이 대표값보다 작은 세트, iii) LPIPS 손실이 대표값보다 큰 세트. 참가자들은 각 세트에서 정체성을 잃지 않으면서 스타일이 충분히 적용된 이미지를 선택하도록 요청받았다. 그림 3. 을 보면, 대부분의 참가자들이 손실이 대표값에 가까운 이미지를 선호하는 경향을 보였다. 이는 LPIPS 기반 대표값이 정체성과 스타일 표현 간의 균형을 반영하는 유의미한 정량적 기준으로 작용함을 시사한다.

3. 결론

본 연구는 사용자 연구를 기반으로 스타일 이미지 생성 과정에서 원본 인물의 정체성 보존과 스타일 표현 간 균형의 기준을 정량적으로 정의하고, 이를 활용한 대규모 데이터셋 구축 방법론을 제시하였다. 다양한 하이퍼파라미터 설정으로 생성된 이미지들에 대해 사용자 연구를 실시하고, LPIPS 값을 분석함으로써 정체성과 스타일 표현 간의 균형을 수치적으로 평가할 수 있는 기준을 마련했다. 이후, 도출된 기준값을 바탕으로 최적 하이퍼파라미터를 역추정하고 추가 사용자 평가를 통해 그 타당성을 검증해, 제안한 방법의 실효성을 확인하였다. 일련의 과정을 통해 본 연구는 사용자 참여 기반의 신뢰성 있는 스타일 이미지 데이터셋 구축 방법을 실증했다. 향후 연구에서는 통계적 신뢰성을 높이기 위해 20명 이상이 참여하는 실험과 5개 이상의 스타일 평가를 진행할 예정이다. 이렇게 생성된 데이터셋은 전문가가 Human-In-The-Loop 방식으로 검수할 계획이며, 최종적으로 스타일 이미지로부터 원본 인물의 정체성을 판별할 수 있는 인공지능 기술 개발에 활용될 수 있을 것으로 기대된다.

참고문헌

[1] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018*, pp. 586-595

XR 환경 내 물리량 연동형 제스처 기반 실시간 특수효과 상호작용 시스템

최종민[0], 장재범, 한도연, 송오영[1]
세종대학교 소프트웨어학과
103solomon@sju.ac.kr, wkdwobum@sju.ac.kr, hando715@gmail.com, oysong@sejong.edu

Real-time Special Effects Interaction System Based on Physical Quantity Linked Gestures in XR Environment

Jong-min Choi[0], Jae-bum Jang, Do-yeon Han, Oh-young Song[1]
Dept. of Software, Sejong University

요약

최근 XR(확장현실) 기술은 하드웨어와 소프트웨어의 발전을 통해 몰입감 있는 사용자 경험을 제공하고 있으며, 고도화된 손 추적 기능을 제공하여 사용자는 컨트롤러 없이도 손의 움직임을 정밀하게 인식하고 상호작용할 수 있게 되었다. 이러한 기술은 XR 환경에서의 행동 언어 기반 인터페이스 개발에 기여하고 있다. 본 논문에서는 손의 속도와 같은 현실 제스처의 물리량을 실시간으로 추정하여 XR 환경 내 특수효과에 직접적으로 반영하는 상호작용 시스템을 제안한다. 시스템은 Unity 엔진 및 XR Interaction Toolkit을 기반으로 구현하고, Meta Quest Pro를 통해 테스트하였으며, 사용자 손 부채질 속도에 따라 발생하는 바람에 영향 받는 Stable Fluids 기반 유체 시뮬레이션과 다층 블록과 손 충돌 시 충격 세기에 따른 물리 반응 시나리오를 구현하여 제안 시스템의 효과적인 상호작용 가능성을 검증하였다.

1. 서론

최근 XR 기술과 손 추적 하드웨어의 발전으로, 사용자의 실제 움직임을 가상 환경에 더 자연스럽게 반영하는 인터페이스 연구가 활발해지고 있다. 기존의 XR 제스처 인터페이스 연구[1]는 주로 손가락 모양과 같은 기호적 제스처 인식에 초점을 두고 있었으며, 실제 손의 속도나 힘 등의 물리량은 상호작용에 직접적으로 반영되지 않았다. 하지만 기호적 제스처만으로는 유체 도메인 내의 국소적인 변화 및 물리 역학적 반응과 같이 정교한 상호작용을 구현하는 데에는 한계가 있다.

본 논문에서는 손의 속도와 같은 물리량을 실시간으로 추정하여 XR 환경의 특수효과 및 객체에 직접적으로 반영하는 시스템을 제안한다.

2. 시스템 설계 및 구현

2.1. 제스처 물리량 추정

본 시스템에서는 사용자의 손 움직임을 정량적으로 분석하기 위해, 각 update 주기마다 손 관절의 위치 데이터를 실시간으로 수집하였다. 본 연구에서는 Meta Quest Pro의 손 추적 기능을 활용하여 손목 및 손가락 관절의 3차원 위치 정보를 받아 사용하였다. 속도는 일정 시간 간격(Δt)마다의 같은 관절의 현재 위치와 이전 위치의 차이(Δx)를 통해 산출한다. 산출된 속도와 객체 및 손의 질량 데이터를 활용하여 물체에 가해지는 충격량을 계산하였다. 계산에 사용된 손의 질량은 성인 남성 기준 실험적 설정으로 400g으로 상정하였다. 이를 통해 손의 운동 방향 및 세기, 객체에 가해지는 충격량을 추정하고, 상호작용이 발생한 특수효과 및 객체에 적용하였다.

2.2. 유체 특수효과 구현

본 연구에서는 유체 기반 특수효과를 시각화하기 위해 GPU에서 실시간으로 Stable Fluids 기반의 유체 시뮬레이션[2,3]을 구현하였다. 시뮬레이션 구현은 Unity 엔진의 셰이더를 활용하여 이루어졌으며, Stable Fluids 알고리즘을 기반으로 다양한 기능을 확장하였다. 주요 연산 단계는 다음과 같다.

1. 유체 대류를 위한 단계(Advection)
2. 외부 힘 적용 단계 (Force kernel)
3. 압력 보정 단계 (Divergence 계산 및 Jacobi 반복)
4. 와류 보강 단계 (Vorticity Confinement)

이 중 외부 힘 적용 단계에서는 사용자의 제스처 물리량(손의 속도 등)이 실시간으로 반영되어 유체 흐름에 직접적인 영향을 미친다. 또한 대류 단계에서는 장애물 내부 및 경계면에서 Free-slip 및 이동 경계 조건을 적용하여, 장애물이 정지해 있거나 움직일 때 유체와의 상호작용이 자연스럽게 재현되도록 구현하였다.

---

[1] 교신저자
* 포스터발표논문, 학부생 주저자 논문임
* 본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2022-00156354), 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업의 지원(IITP-2025-RS-2023-00254529), 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 지원(2024-0-00037)을 받아 수행되었음.

3. 시스템 구현 및 결과

3.1. 손과 유체 특수효과의 상호작용

본 시스템에서는 사용자가 손으로 부채질을 할 경우, 손의 이동 방향과 속도를 실시간으로 추정하여 해당 방향으로 바람 효과를 발생시킨다. 특히, 손의 빠르기가 빠를수록 유체에 가해지는 힘의 크기가 커져, 강한 손동작일수록 유체(연기)의 움직임이 더욱 크게 나타난다. 이 바람 효과는 Stable Fluids 기반으로 시뮬레이션 된 연기 필드에 직접적으로 적용되어, 손의 움직임에 따라 연기의 방향, 속도, 확산 패턴이 실시간으로 변한다. 그림 1에서 사용자가 손을 빠르고 크게 휘두르면 연기가 해당 방향으로 강하게 밀려나거나 확산되며, 확산 과정은 장애물의 영향을 받는다. 느린 손동작에는 그에 비례하여 약한 바람과 미세한 연기 변화만 나타난다. 이를 통해 사용자는 손의 속도와 방향만으로도 직관적으로 가상 유체(연기)의 거동을 조작할 수 있다.

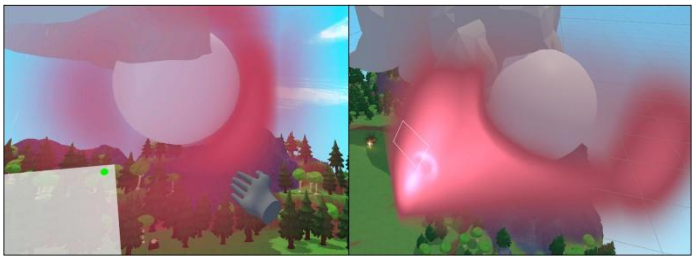


그림 1: 제스처의 물리량을 반영한 유체와의 상호작용

3.2 충격량 기반 객체 상호작용

본 시스템에서는 손이 쌓여 있는 박스 구조물의 중간 또는 아래쪽 박스에 충돌할 때, 손의 속도와 가상 질량을 곱한 충격량을 실시간으로 계산한다. 그림2의 첫번째 사진과 같이 사용자가 충분히 강한 힘(충격량)으로 중간 또는 아래 박스를 치면, 해당 박스만 구조물에서 빠져나가거나 튕겨 나가고, 그 위에 쌓여 있던 박스들은 중력에 의해 자연스럽게 아래로 낙하한다. 반대로, 손의 힘이 약한 경우에는 그림2의 두번째 사진처럼 충격을 받은 박스와 함께 전체 구조물이 무너지거나 흐트러지는 현상이 나타난다. 또한, 그림3에서 충격량에 의해 가속된 공은 변화된 속도를 기반으로 주변 유체와 상호작용하며, 이에 따라 유체 내에 난류나 흐름 변화가 발생한다. 충격량이 작을 경우, 공이 낮은 속도로 움직여 그림3의 첫번째 사진처럼 유체가 넓게 퍼지는 모습을 볼 수 있고, 충격량이 클 경우 공이 빠른 속도로 움직이게 되어 두번째 사진처럼 유체가 안쪽으로 말려들어가는 것을 볼 수 있다. 이와 같은 구현을 통해 사용자는 손의 힘 조절에 따라 충격량을 전달함 으로서 유체 고체 모두에서 실제 물리현상과 유사한 효과를 얻을 수 있다. 이 기능은 XR 환경 내에서 보다 직관적인 객체 상호작용과 현실적인 물리 반응을 제공한다.

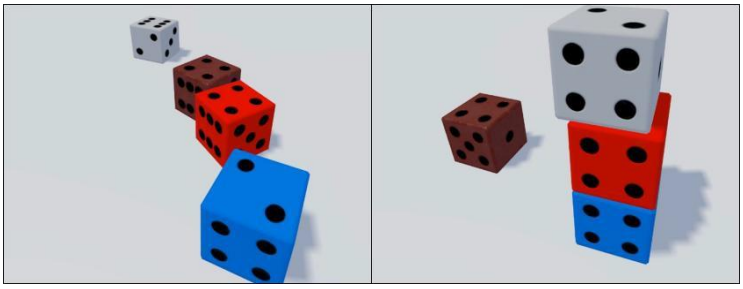


그림 2: 충격량에 기반한 다층 오브젝트 상호작용

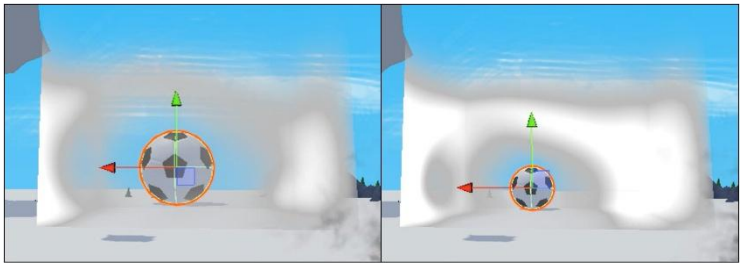


그림 3: 충격량에 의해 가속된 공의 속도 변화와 유체와의 상호작용 과정

4. 결론

본 논문에서는 손의 실제 속도와 충격량 등 현실적인 제스처 물리량을 XR 환경의 특수효과 및 객체에 직접 적용하는 상호작용 시스템을 제안하였다. 사용자가 손으로 부채질하는 방향과 속도에 따라 가상 유체의 움직임이 직관적으로 변화하며, 쌓여 있는 다층 구조물의 특정 박스를 강하게 치는 경우 해당 박스만 빠져나가고 위에 쌓인 물체는 자연스럽게 낙하하는 등, 현실적인 상호작용을 재현하였다. 본 연구는 기존의 기호적 제스처만을 사용하는 기존 XR 인터페이스의 한계를 극복하고, 정량적 물리량을 기반으로 한 물리 역학적 상호작용이 XR 환경에서 효과적으로 구현 가능함을 확인하였다. 향후 연구에서는 손의 가속도, 각속도 등 다양한 물리량의 적용, 더욱 복잡한 객체·환경에서의 상호작용이 필요한 다양한 사용자 시나리오(예: 교육용 콘텐츠, 게임 인터렉션 등)에 맞춰 제스처-물리량 매핑 방식을 최적화하는 후속 연구를 통해 시스템의 적용 범위와 현실감을 더욱 높일 계획이다.

참고문헌

[1] 최종민; 김민채; 송오영. XR 환경 특수효과 제어를 위한 멀티모달 인터페이스. 한국컴퓨터그래픽스학회 학술대회, 2024, 153-154.

[2] Stam, Jos. "Real-time fluid dynamics for games. "Proceedings of the game developer conference. Vol. 18. No. 11. 2003.

[3] Fedkiw, Ronald, Jos Stam, and Henrik Wann Jensen. "Visual simulation of smoke." Proceedings of the 28th annual conference on Computer graphics and interactive techniques. 2001.

인간-로봇 상호작용을 위한 비언어적 행동 시뮬레이션 구현*

김동민†,0, 장기현†, 이강훈
광운대학교 소프트웨어학부
eastals@kw.ac.kr, jgh0411@naver.com, kang@kw.ac.kr

Implementation of Nonverbal Behavior Simulation for Human-Robot Interaction

Dong Min Kim†0, Gi Hyeon Jang†, Kang Hoon Lee
Department of Software, Kwangwoon University

요약

본 연구에서는 향후 사람과의 자연스러운 상호작용 연구를 위한 기반으로, 시각 및 음성 정보를 활용하여 사용자를 인식하고 추적하는 로봇 시뮬레이션 시스템을 구현하였다. 이를 위해 3D 게임 엔진인 유니티(Unity)와 로봇 운영체제(robot operating system, ROS)를 연동하고, 실제 하드웨어 제어 방식과 유사한 물리 제어 기능을 갖춘 통합 시뮬레이션 환경을 구축하였다. 유니티는 다양한 가상 환경을 구성할 수 있는 유연성을 제공하며, ROS는 MoveIt 프레임워크를 통해 로봇의 움직임을 계획하고 제어한다. 제안된 시스템은 주변에서 발생하는 소리의 방향을 추정한 뒤, 시각 정보를 통해 얼굴을 인식하고 로봇이 사용자를 바라보도록 자세를 조정한다. 본 시뮬레이션은 실제 하드웨어 없이도 신뢰도 높은 실험이 가능하며, 향후 인간-로봇 상호작용 연구에 있어 유용한 실험 플랫폼으로 활용될 수 있다.

1. 서론

오늘날 키오스크나 챗봇 등 다양한 형태의 인간-기계 상호작용이 일상화되면서, 로봇과의 직접적인 상호작용을 다루는 HRI(human-robot interaction) 연구도 활발히 진행되고 있다. 특히, 인공지능을 로봇에 내장하는 임바디드 AI 기술이 발전함에 따라, 음성 명령을 넘어 시선, 자세, 제스처 등 비언어적 행동이 상호작용의 질을 높이는 핵심 요소로 주목받고 있다. Mavridis[1]는 비언어적 단서가 인간-로봇 간 의사소통에서도 중요한 역할을 한다고 지적하였으며, Rosenthal-von der Pütten[2]은 로봇이 인간과 유사한 비언어적 표현을 수행할 때 사용자의 공감 및 사회적 유대감이 증진된다고 보고하였다. 그러나 이러한 실험을 실제 로봇으로 수행하는 데는 비용, 안전성, 환경 제약 등 현실적 한계가 따른다. 이에 본 연구에서는 비언어적 행동 기반의 인간-로봇 상호작용을 가상 환경에서 실험할 수 있는 시뮬레이션 시스템을 구현하였다. 본 시뮬레이션은 유연한 가상 환경 구성이 가능한 3D 게임 엔진인 유니티와 로봇 제어에 널리 사용되는 ROS를 연동하여, 실제 하드웨어 없이도 신뢰도 높은 상호작용 실험이 가능하도록 설계되었다[3].

* 포스터 발표논문
* † 동일하게 기여함
* 학부생 주저자 논문임
* 본 연구는 카카오페이에서 후원하는 '디지털교육격차해소를 위한 IT교육지원' 발전기금으로 수행되었음.

2. 시뮬레이션 구조

2.1. 로봇 시뮬레이션

유니티 상에서 실제 하드웨어를 가진 로봇을 시뮬레이션하기 위해, ROS를 활용하여 하드웨어 간의 통신과 제어를 수행하였다. MoveIt을 이용하여 역기구학 계산과 동작 계획을 수행하고, ROS 기반 제어 노드를 통해 로봇을 제어하였다[4]. 상용 로봇 모델을 기반으로 시뮬레이션을 구성하고, 제조사에서 제공하는 관성 모멘트, 관절 제한각도 등의 물리 파라미터를 반영하여 실제 동작과 유사한 물리 특성을 갖도록 구성하였다.

2.2. 시각과 소리를 이용한 사람 추적

소리 기반 추적을 위해 가상 마이크 배열을 구성하고, 소리의 도착 시간 차(time difference of arrival, TDOA)를 이용해 음원의 방향을 추정하였다. 소리가 감지되면 로봇은 해당 방향으로 회전하고, 이후 시각 정보를 활용한 얼굴 추적 단계로 전환된다. 유니티의 카메라 컴포넌트를 통해 로봇의 시각 정보를 획득하며, 딥러닝 기반 얼굴 검출 모델을 통해 사용자의 얼굴을 인식한다. 로봇은 얼굴이 화면 중앙에 위치하도록 회전 및 이동하고, 얼굴 크기가 사전에 정의된 기준에 도달할 때까지 상대 거리 조정을 수행한다.

2.3. 유니티-ROS 통합 시스템

유니티와 ROS 간 양방향 통신을 TCP를 통해 구현하여 ROS에서 수행하는 로봇 제어와 유니티에서 실행되는

사용자 인식 및 환경 시뮬레이션을 실시간으로 연동하는 통합 시스템을 구성하였다. 유니티에서 계산된 얼굴 위치와 음원 방향 정보를 ROS로 전달하면, MoveIt이 이를 기반으로 경로를 계획하고 로봇 제어를 수행한다. ROS에서 갱신된 로봇 상태는 다시 유니티로 전송되어 시각적 피드백과 시뮬레이션 환경을 동기화한다. 이를 통해 실제와 유사한 상호작용을 가상 환경 내에서 구현할 수 있다.

3. 실험 결과

시뮬레이션에서는 탁자 위에 고정된 로봇이 시각 및 청각 정보를 이용해 사용자를 인식하고 추적하는 시나리오를 구현하였다. 초기 상태에서 로봇은 마이크와 카메라를 통해 주위 환경을 감지하며, 사용자에 대한 신호가 없을 경우 좌우로 회전하며 주변을 탐색한다. 그림 1은 로봇이 배치된 초기 시뮬레이션 환경을 보여준다. 청각 입력이 감지되면 TDOA 기반으로 추정한 음원 방향을 따라 로봇이 회전한 뒤, 시각 추적 단계로 전환된다. 그림 2와 그림3은 시간 순서에 따라 로봇의 동작 변화를 시각화한 것으로, 각 사진은 유니티 씬 뷰(상)와 로봇에 부착된 카메라 화면(하)으로 구성되어 있다. 그림2는 로봇이 음원 방향으로 회전하는 과정을, 그림 3은 얼굴 탐지 및 추적 과정을 단계별로 보여준다. 로봇은 음원 방향으로 회전한 뒤, 카메라 화면 내 얼굴이 화면 중앙에 오도록 자세를 조정하고, 사용자가 이동하면 다시 얼굴을 인식해 새로운 방향으로 회전하여 추적을 이어간다.

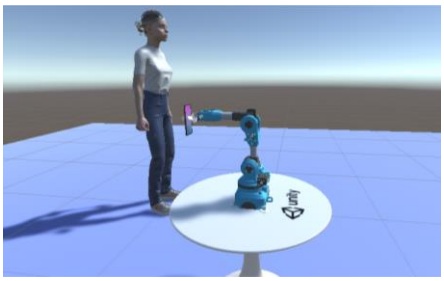


그림 1: 시뮬레이션 환경

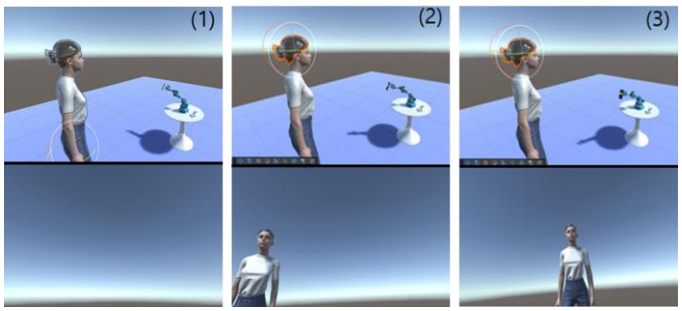


그림 2: 소리의 도착 시간 차를 이용한 음원 발생 위치 추적 (왼쪽부터 시간순으로)

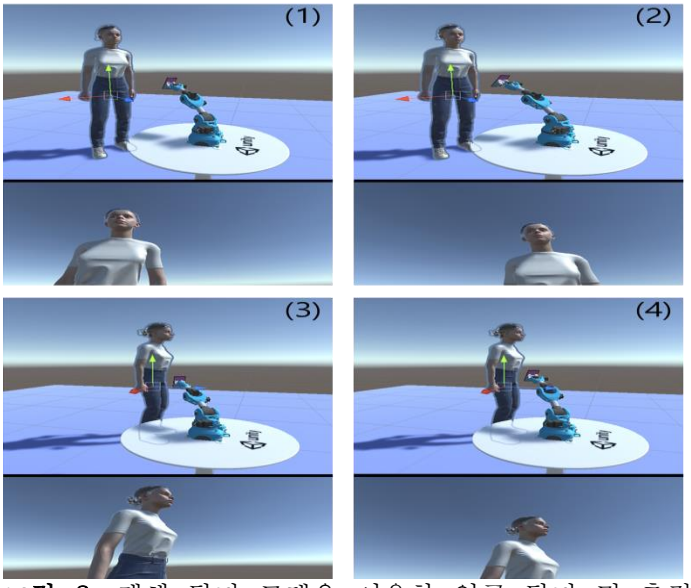


그림 3: 객체 탐지 모델을 이용한 얼굴 탐지 및 추적 1) 초기 상태, 2) 얼굴을 인식하고 화면 중앙에 위치하도록 로봇을 이동, 3) 사용자가 새로운 위치로 이동, 4) 새로운 위치에서 발견된 얼굴이 중앙에 위치하도록 이동

4. 결론 및 향후 계획

본 논문에서는 ROS와 유니티를 연동하여, 실제 하드웨어와 유사한 방식으로 제어 가능한 로봇 시뮬레이션을 구축하였다. 유니티의 유연한 환경 구성 능력과 ROS의 물리 기반 제어 기능을 통합함으로써, 시각 및 음성 정보를 기반으로 사용자를 인식하고 추적하며 마주 보는 상호작용 시나리오를 효과적으로 재현할 수 있었다. 이러한 통합 시뮬레이션 구조는 실제 환경을 대체할 수 있는 실험 플랫폼을 제공하며, 향후 인간-로봇 상호작용 연구에 활용될 수 있다.

참고문헌

[1] N. Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63: 22-35, 2015.
[2] A. M. Rosenthal-von der Pütten, N. C. Krämer, et al. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5: 17-34, 2013.
[3] S. Macenski, T. Foote, et al. Robot operating system 2: Design, architecture, and uses in the wild. *Science robotics*, 7.66: eabm6074, 2022.
[4] D. Coleman, I. Sucan, et al. Reducing the Barrier to Entry of Complex Robotic Software: a MoveIt! Case Study, *Journal of Software Engineering for Robotics*, 5(1): 3–16, 2014.

3D 가우시안 스플래팅을 위한 암묵적 신경망 표현

김승겸[01], 정인재[2], 김아름[3], 유용재[1], 윤석민[1]
[1]한양대학교 인공지능융합학과, [2]한양대학교 ERICA 수리데이터사이언스학과, [3]한양대학교 ERICA 인공지능학과
{skkim3533, youjae0920, dkfma0817, yongjaeyoo, sukminyun}@hanyang.ac.kr

Implicit Neural Representations for 3D Gaussian Splatting

Seung-gyeom Kim[01], Injae Jung[2], Areum Kim[3], Yongjae Yoo[1], Sukmin Yun[1]
[1]Dept. of Applied Artificial Intelligence, Hanyang University
[2]Dept.of Mathematical Data Science, Hanyang University ERICA
[3]Dept. of Artificial Intelligence, Hanyang University ERICA

요약

본 논문에서는 3D Gaussian Splatting(3DGS)의 확장성 한계를 해결하기 위해, 가우시안 파라미터를 3D 공간 좌표로부터 예측하는 암묵적 신경망 표현(Implicit Neural Representation) 기반의 학습 프레임워크를 제안한다. 제안된 방법은 매 반복마다 가우시안의 일부 좌표만을 사용해 파라미터를 예측하는 신경망을 학습하며, 렌더링은 전체 가우시안에 대해 수행된다. 이를 통해 기존 방식처럼 모든 가우시안을 한 번에 메모리에 올릴 필요 없이 미니배치 학습이 가능해지고, 역전파 시 GPU 메모리 사용량이 절감된다. 또한, 기존의 NeRF 기반 암묵 표현 기법들과 달리, 본 방법은 생성된 가우시안 파라미터를 이용해 기존 3DGS와 유사한 속도로 빠르게 렌더링할 수 있다는 장점을 가진다. 실험 결과, Tanks and Temples 데이터셋에서 기존 3DGS 대비 더 우수한 시각적 품질을 보여주며, 본 방식이 계산 효율성과 품질 면에서 모두 효과적임을 입증하였다.

1. 서론

최근 3D Gaussian Splatting(3DGS)[1]은 빠른 추론 속도와 뛰어난 렌더링 품질로 Novel view synthesis 분야에서 주목을 받고 있다. 3DGS는 장면을 수많은 3D 가우시안으로 표현한 뒤, 이를 2D 공간에 투영하고 래스터화(rasterization)를 통해 이미지를 생성한다. 이 과정에서 사용되는 미분 가능한 래스터화 기법은 픽셀 단위로 렌더링 손실을 계산할 수 있도록 하며, 이를 통해 가우시안 파라미터를 직접 최적화할 수 있다.

그러나 수백만에서 수천만 개의 가우시안을 명시적으로 최적화해야 하므로, GPU 메모리 사용량이 크고 단일 GPU 환경에서는 확장성에 대한 한계가 존재한다. 확장성 문제를 해결하기 위해, 가우시안들을 여러 GPU에 분산시켜 병렬로 렌더링 및 최적화를 수행하는 Mulit-GPU 학습 방식이 제안되었다.

* 포스터 발표논문
* 본 논문은 요약논문 (Extended Abstract)으로서, 본 논문의 원본 논문은 현재 타 학술대회(논문지)에 제출중
* 본 연구는 2023년 과학기술정보통신부 기초연구사업(우수신진연구, RS-2023-00212914) 및 중소기업기술정보진흥원 중소기업기술 혁신 개발사업 (RS-2025-02304702)의 지원을 받아 수행되었다.

Multi-GPU 학습 방식은 메모리 사용의 효율성을 개선할 수 있지만, GPU 간 통신 오버헤드가 추가로 발생하기 때문에 GPU 수가 많아질수록 학습 속도가 저하된다. 따라서 학습 속도저하를 방지하기 위해, 메모리가 제한된 단일 GPU 환경에서도 확장 가능한 학습방식이 필요하다.

본 논문에서는 단일 GPU 환경에서도 확장 가능한 새로운 학습 방법론을 제안한다. 암묵적 신경망 표현을 활용하여 3D 공간좌표로부터 가우시안 파라미터를 생성하며, 매 학습 단계에서 가우시안 일부만 선택적으로 업데이트함으로써 GPU 메모리 사용량을 줄이고 학습 효율을 향상시킨다.

2. 방법

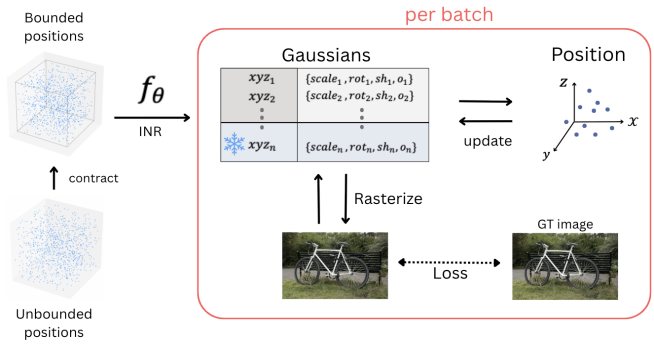


그림 1: 방법론 개요

암묵적 신경망 표현을 통해 3DGS에서 사용되는 가우시안 파라미터를 생성하고, 이를 기반으로 확장 가능한 네트워크 구성과 배치 단위 가우시안 업데이트를 통한 최적화 방법을 제안한다(그림 1 참조).

2.1 네트워크

본 연구에서는 암묵적 신경망 표현을 기반으로 3D 가우시안의 주요 파라미터들을 공간 위치로부터 직접 예측하는 방식을 사용한다. 기존 3DGS 방식에서는 각 가우시안 파라미터(크기, 회전, 색상 계수, 불투명도)를 개별적으로 직접 최적화했으나, 본 연구에서는 Instant-NGP(INGP)[2] 구조를 활용하여 3D 위치 정보를 고차원 임베딩으로 변환하고, 이를 shared implicit representation으로 삼아 여러 신경망을 통해 각 파라미터를 생성 및 업데이트한다. 각 파라미터는 전용 MLP

디코더를 통해 분리되어 예측되며, 안정적인 학습을 위해 디코더별로 적절한 초기 가중치 값이 설정된다.

2.2 배치 단위 가우시안 업데이트 기법

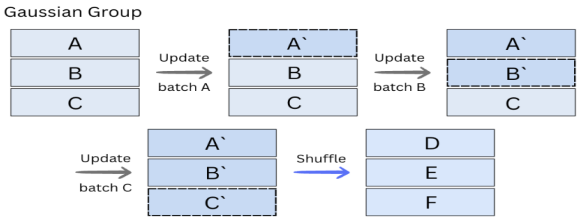


그림 2: 배치별 가우시안 업데이트

모든 가우시안 파라미터를 매 반복마다 업데이트하는 기존 방식은 GPU 메모리 사용량이 매우 높고, 대규모 장면 학습에 비효율적이다. 이에 본 연구는 학습 시 매 반복(iteration)마다 전체 가우시안 중 일부만 선택적으로 업데이트하는 배치 단위 가우시안 업데이트 기법을 도입하였다(그림 2 참조). 모든 배치가 한 차례씩 업데이트되면, 가우시안들을 무작위로 섞어(shuffle) 새로운 배치를 구성한다. 이 방식은 전체 파라미터 공간에 대한 표현력을 유지하면서도 메모리 사용량을 크게 절감할 수 있으며, 단일 GPU 환경에서도 수많은 가우시안을 효과적으로 학습할 수 있는 확장성을 제공한다. 이를 통해 학습 속도와 자원 효율성 모두를 향상시킬 수 있다.

3. 실험

실험 환경은 24GB 메모리를 탑재한 RTX 4090 GPU에서 PyTorch 2.1 및 CUDA 11.8 환경으로 수행되었으며, 대규모 가우시안 학습이 필요한 경우에는 NVIDIA A100 80GB GPU를 사용하였다. 3.1절에서는 제안하는 암묵적 신경망 표현 기반 방식이 기존 3DGS 방식에 비해 얼마나 우수한지를 평가하였으며, 비교 실험을 위해 초기 가우시안 위치를 기존 3DGS 방식으로부터 얻은 결과를 활용하였다. 3.2절에서는 매우 많은 수의 가우시안을 효율적으로 학습할 수 있는지를 검증한다.

3.1 비교 실험

Methods	Train			Truck			Avg		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
INGP	20.17	0.666	0.386	23.26	0.779	0.274	21.715	0.7225	0.330
3DGS	21.10	0.816	0.218	25.19	0.882	0.148	23.145	0.849	0.183
Ours	22.89	0.823	0.205	25.67	0.881	0.156	24.28	0.852	0.180

표 1: INGP, 3DGS, Ours 비교 실험 결과

(a) Ours (b) 3DGS

그림 3: Tanks&Temples 결과의 정량적 및 정성적 비교

Tanks & Temples 데이터셋[3]에서 제안한 프레임워크는 Train 장면에서 22.89, Truck 장면에서 25.67 PSNR 점수를 기록하며 뛰어난 성능을 보였다. 또한, 기존 방법인 INGP와 3DGS과 비교했을 때, 각각의 장면뿐만 아니라 평균 PSNR에서도 크게 향상된 결과를 나타냈다. 특히 평균 PSNR은 24.28로 3DGS의 23.15, INGP의 21.72에 비해 우수한 성능을 확인할 수 있다. 더불어 SSIM과 LPIPS 지표에서도 제안한 방법이 가장 우수한 수치를 기록하여, 시각적 품질과 유사도 측면에서도 뛰어난 성능을 입증하였다(표 1, 그림 3 참조).

3.2 많은 수의 가우시안을 훈련시키는 실험

Method / Batch Size	PSNR↑	SSIM↑	Memory	Device
3DGS	25.18	0.7615	26GB	A100
Ours (0.5M)	23.15	0.682	21GB	RTX 4090
Ours (1M)	24.18	0.712	23GB	
Ours (2M)	25.01	0.747	26GB	
Ours (3M)	25.09	0.754	32GB	A100
Ours (4M)	25.11	0.756	36GB	

표 2: Bicycle 데이터셋에서의 대규모 학습

Mip-NeRF 360의 Bicycle 데이터셋을 사용해 1,400만 개의 가우시안을 A100 (80GB) GPU에서 생성 및 학습한 결과 (표 2 참조), 기존 3DGS는 PSNR 25.18을 달성했으나, 동일 규모를 RTX 4090 (24GB) 환경에서 학습하는 것은 메모리 한계로 인해 불가능 하다. 반면, 본 연구의 프레임워크는 RTX 4090에서도 PSNR 24.18에 도달하며 3DGS와 유사한 수준의 성능을 달성하였다. 또한, 배치 크기가 증가함에 따라 메모리 사용량과 학습 시간이 함께 증가하는 명확한 트레이드오프가 관찰되었으며, 이는 제안한 방법이 제한된 하드웨어 환경에서도 실용적인 대안이 될 수 있음을 시사한다. 다만, 배치 사이즈가 작다보니 성능 향상에는 제약이 있었다.

4. 결론

본 연구에서는 효율적이고 확장 가능한 3D Gaussian Splatting 학습 기법을 제안한다. 기존 3DGS의 GPU 메모리 한계를 해결하기 위해, 가우시안 파라미터를 공간 좌표로부터 공유된 암묵적 표현으로 학습하고, 학습 시 미니배치 기반으로 처리하여 메모리 사용량을 효과적으로 절감한다. 이 방식은 GPU 메모리 효율성을 크게 개선하면서도 기존 3DGS와 유사한 수준의 빠른 렌더링 속도와 높은 시각적 품질을 유지한다. 특히 가우시안 수가 증가하는 대규모 장면에서도 안정적인 학습과 렌더링 성능을 보인다. 본 연구는 가우시안 스플래팅 파라미터에 대한 암묵적 표현 기반 접근법의 활용 가능성을 보여주었으며, 향후 3D 및 4D 장면 렌더링의 확장성과 효율성 향상에 기여할 수 있을 것으로 기대한다.

참고문헌

[1] Kerbl, Bernhard, et al. "3d gaussian splatting for real-time radiance field rendering." ACM transactions on graphics, 42.4 ,139:1-139:14, 2023
[2] Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." ACM transactions on graphics (TOG), 41.4, 1-15, 2022.
[3] Knapitsch, Arno, et al. "Tanks and temples: Benchmarking large-scale scene reconstruction." ACM Transactions on Graphics (ToG), 36.4, 1-13, 2017.

애플 비전 프로를 활용한 XR 기반 물리 교육 시뮬레이션 개발*

정구현[0], 장윤석, 이규민, 한도연, 송오영[1]
세종대학교
{wrg267499, ttttwww10, lewis719}@naver.com, hando715@gmail.com, oysong@sejong.edu

XR based physics education leveraging apple vision pro

Goohyun Jeong[0], Yunseok Jang, Kyumin Lee, Do-yeon Han, Oh-young Song[1]
Sejong University

요약

현재 교육용 물리 시뮬레이션은 컴퓨터나 스마트폰을 활용하여 2차원 화면상에서 진행된다. 이러한 2차원 자료 역시 학습에 도움이 되지만, 몰입감에는 한계가 있다. 최근 애플 비전 프로와 메타 퀘스트3 등의 XR 장비가 개발되었으며, 이러한 장비는 사용자로 하여금 XR환경을 통해 보다 큰 몰입감을 경험하게 한다. 본 연구에서는 다양한 XR 장비 중 애플 비전 프로를 사용해 시뮬레이션을 진행하였다. 현재 애플 비전 프로에 있는 컨텐츠들은 물리 현상을 물리적으로 정확하게 재현하지 않거나 시뮬레이션 제어에 대한 자유도가 낮은 경우가 많다. 이를 해결하기 위해 애플 비전 프로에서 쉽게 제어 가능한 시뮬레이션 컨텐츠 개발을 통해 물리 교육의 효율성을 향상시키고자 하였으며, 설문조사를 통해 웹 기반 시뮬레이션 보다 XR 물리 시뮬레이션 환경이 물리 교육에 더 효과적임을 알 수 있었다.

1. 서론

XR 콘텐츠는 다양한 분야의 교육에 활용되고 있으며, 물리 교육을 위한 XR 콘텐츠는 보다 효과적인 학습을 가능하게 한다[1]. 하지만 비전 프로의 물리 관련 컨텐츠에서 태양계의 공전과 자전의 경우, 실시간 물리 역학 계산을 거치지 않고 사전에 정해진 궤도를 따라 시각화되어 실험 변수의 실시간 조정을 통한 체험이 어려워 학습의 효과가 감소한다. 이를 위해 본 연구에서는 물리 역학을 실시간으로 계산하여 시뮬레이션 함으로서 실험 변수 조정을 통해 XR 환경에서의 물리 실험에 대한 몰입감을 높이고 교육의 효과를 향상시키고자 하였다.

* 구두(포스터) 발표논문
* [1]교신저자
* 본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2022-00156354), 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업의 지원(IITP-2025-RS-2023-00254529), 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 지원(2024-0-00037)을 받아 수행되었음.

2. 물리 교육과 소프트웨어 활용

2.1. 기술의 발전과 물리 교육

전통적인 물리 교육은 이론과 실제 실험을 기반으로 진행되어 왔다. 하지만 실제 실험은 비용 및 안전성 등의 문제로 많은 제약이 존재한다. 이러한 제약을 해결하기 위해 현대에는 영상, XR 등의 기술을 물리 교육에 활용하고 있다. 영상을 활용하는 경우 사전에 녹화된 실험 영상에 CG를 추가하거나 컴퓨터 시뮬레이션을 통해 계산하여 출력된 영상을 사용한다. XR을 활용하는 경우 실제 환경 및 객체 위에 물리 현상이 적용되는 모습을 증강하여 시각화 하거나 가상의 공간에서 다양한 실험 환경 및 도구와의 상호작용으로 물리 현상을 체험하며 특히, 위험하거나 접근이 어려운 환경에서의 실험을 가능하게 한다. 이 외에도 인공지능을 활용[2]하는 등 기술의 발전으로 물리 교육을 위한 도구가 다양해지고 있다.

2.2. 물리 교육과 관련된 기존 소프트웨어

2024년에 출시한 애플의 비전 프로는 높은 해상도와 낮은 지연으로 더 큰 몰입감을 제공한다. 애플 비전 프로의 컨텐츠는 감상에 초점을 맞춘 컨텐츠가 주를 이룬다. 하지만 물리 교육은 체험이 가능할 때 그 효과가 극대화되기 때문에 애플 비전 프로의 장점을 활용하기 위해서는 사용자와의 상호작용 기능이 필요하다. solAR[3]과 같은 기존에 배포된 시뮬레이션 관련 프로그램은 위성의 크기와 시뮬레이션의 진행속도는 제어할 수 있으나 다양한 설정의 체험에는 한계가 있다.

3. XR 기반 물리 교육 시뮬레이션 소프트웨어

3.1. 활용 프레임워크 및 개발 방법

개발 시 주요모듈로 Swift의 RealityKit을 활용했으며 오브젝트들을 클래스로 관리해 개발의 효율성을 높였다. 같은 프로그램에서 여러가지 예제를 시뮬레이션 할 수 있도록 하였다. 사용자는 UI 버튼을 통해 구현되어 있는 예제를 시뮬레이션 할 수 있다.

3.2. 물리 시뮬레이션 응용 사례

뉴턴의 운동 법칙과 만유인력의 법칙을 기반으로 천체 운동을 이해하는 교육 프로그램을 개발했다. 천체의 위치와 속도가 실시간으로 계산될수 있도록 시뮬레이션 시간간격을 적절히 조절하였다. 태양계를 구성하는 천체들은 실제 질량과 거리를 사용했으며 효과적인 시각화를 위해서 렌더링 시 일부 오브젝트의 위치와 크기를 수정하였다.

그림 1: 태양, 지구

UI를 통해 예제를 전환할 수 있고, 계산된 내용(속력이나 힘 등)을 확인할 수 있다. 또한 그림2와 같은 UI를 통해 특정 예제에서는 질량이나 시간 간격을 조정해서 그림3과 같이 달라지는 운동을 확인할 수 있다. 또한 버튼을 통해 카메라의 시점을 이동시키는 등의 상호작용이 가능하다.

그림 2: UI창

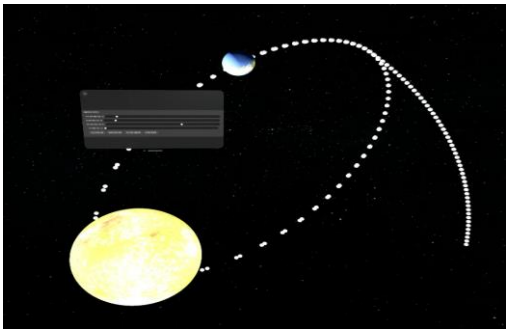


그림 3: 태양의 질량을 증가시켰을 때

4. 설문조사 결과

설문조사 응답자들 중 XR컨텐츠를 경험해본 사람은 5명이며, 3명은 XR경험를 경험해 본 적이 없다. 설문조사는 5점 척도(1점:전혀 아니다, 2점:아니다, 3점:보통이다, 4점:그렇다, 5점:매우 그렇다)로 평가받았으며, 각 문항은 몰입도가 높았는지, 조작이 편했는지, 학습 내용에 대한 흥미가 증가하였는지에 대해 물었다. XR 시뮬레이션에서 몰입도는 5점이 6명, 조작도는 2점과 3점이 3명, 흥미도는 5점이 5명으로 제일 많다. 웹 기반 시뮬레이션에서 몰입도는 2점이 4명, 조작도는 5점이 3명, 흥미도는 1점과 2점이 3명으로 제일 많다. 표1과 표2에서 확인할 수 있듯이 XR 시뮬레이션에서 조작도가 비교적 낮은 점수를 보여주는데 이는 아직 XR 기기가 보급되지 않았단 점에서 더 낯선 조작 방법이기 때문인 것으로 보인다.

분류	1점	2점	3점	4점	5점
몰입도	0	0	0	2	6
조작도	0	3	3	1	1
흥미도	0	0	1	2	5

표 1: XR 물리 시뮬레이션 평가

분류	1점	2점	3점	4점	5점
몰입도	2	4	2	0	0
조작도	1	0	2	2	3
흥미도	3	3	1	1	0

표 2: 웹 기반 물리 시뮬레이션 평가

본 연구에서는 XR 장비 중 하나인 애플 비전 프로에서 실시간 물리 역학 계산을 통한 물리 현상 시뮬레이션 콘텐츠를 개발하였다. 기존 XR 물리 콘텐츠와 비교하여 실시간으로 물리 역학을 계산함으로 사용자가 UI를 통해 실험 변수를 조정하는 것이 가능하다. 이는 사용자에게 보다 큰 몰입감을 제공하고 학습에 대한 흥미를 높이며 교육 효과를 증진시킨다. 향후 사용자 상호작용의 다양화 및 직관성 강화(예: 자연스러운 손 제스처 인식 및 햅틱 피드백 연동), 시뮬레이션 콘텐츠의 확장 및 심화(다양한 물리 분야 예제 추가 및 심화 학습 시나리오 개발)을 통해 학습자가 물리 현상을 깊이 이해하고 능동적으로 탐구할 수 있는 강력한 교육 도구로 개발하고자 한다.

참고문헌

[1] J.F.Villada Castillo, L.Bohorquez Santiago and S.Martínez García. Optimization of Physics Learning Through Immersive Virtual Reality: A Study on the Efficacy of Serious Games, *Applied Sciences*, 15(6):3405. 2025.

[2] A.Gunturu, Y.Wen, N.Zhang, J.Thundathil, R.H.Kazi and R.Suzuki, Augmented Physics: Creating Interactive and Embedded Physics Simulations from Static Textbook Diagrams, *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24),* Article 144, pp.1-12, New York, NY, Association for Computing Machinery.

[3]https://apps.apple.com/us/app/solar-solar-system-in-ar/id1286558019?platform=appleVisionPro

포토그래메트리 기반 저비용 이동형 3D 손 스캔 시스템*

최재효[0,1], 조혜성[1], 박평화[1], 임재호[2], 한다성[1]
[1]한동대학교, [2]덱스터 스튜디오
{pickpie, 22000728, 22473003}@handong.ac.kr, jaeho.im@dexterstudios.com, dshan@handong.edu

A Low-Cost Portable 3D Hand Scanning System Based on Photogrammetry

Jaehyo Choi[0,1], Hyeseong Choi[1], Pyeonghwa Park[1], Jaeho Im[2], Daseong Han[1]
[1]Handong Global University, [2]Dexter Studios

요약

본 연구는 저비용 이동형 3D 손 스캔 시스템을 제안한다. 기존의 상용 전신 또는 얼굴 스캐너는 일반적으로 수십에서 수백 대의 고급 카메라에 의존하며, 고정된 스튜디오 환경에서의 작동이 필요하고 장비 설치 및 운용을 위해 높은 수준의 전문성이 요구된다. 본 시스템은 라즈베리 파이와 카메라 모듈, 모듈형 프레임을 기반으로 구성되어 이동성과 확장성을 확보하였으며, 손 스캔에 적합한 구조로 설계되었다. 포토그래메트리 기법을 통해 촬영 후 3D 메시로 복원하였으며 선행 연구 대비 적은 장비로도 효과적인 결과를 도출하였다.

1. 서론

3D 스캐닝 기술은 디지털 콘텐츠, 의료, 공학, 문화유산 보존 등 다양한 분야에서 사물의 정밀한 디지털 복제를 가능하게 하며 핵심적인 역할을 한다. 특히 메타버스 및 CG/VFX 산업의 성장과 함께, 인간의 손처럼 복잡한 해부학적 구조를 고해상도로 재구성하는 기술의 수요가 증가하고 있다[1]. 그러나 고정밀 포토그래메트리 기반 상용 스캐너는 수천만 원 이상의 비용과 고정식 환경, 수십~수백 대의 고성능 카메라 및 고도의 운용 기술을 요구해 소규모 연구자나 기관에게는 진입 장벽이 크다. 근래에는 저비용 포토그래메트리 기반 손 스캐너에 대한 연구도 일부 진행된 바 있다. 라즈베리 파이 50대와 카메라 모듈을 이용해 손의 3D 형상을 캡처하는 시스템을 개발하였으나, 큰 부피에 고정적인 형태, 손에 최적화되지 않은 카메라 배치 등 개선점이 존재했다[2]. 본 연구는 라즈베리 파이와 카메라 모듈을 활용하여, 이동 가능한 저비용 소형 3D 손 스캐너를 제안한다. 본 논문의 남은 절에서는 시스템 구현을 위한 하드웨어 설계, 카메라 배치 및 테스트, 시스템 구조 및 데이터 처리 과정 등을 구체적으로 설명한다.

* 포스터 발표논문, 학부생 주저자 논문임
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회 (논문지)에 제출 준비중임.
* This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00229451, Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms).

2. 시스템 개요

고안된 3D 손 스캐너 시스템은 저비용, 고정밀, 모듈형 확장성이라는 세 가지 주요 목표를 바탕으로 설계되었다. 기존 상용 시스템이 고가의 장비, 고정된 스튜디오 환경, 숙련된 전문 인력에 의존하는 한계를 극복하기 위해, 라즈베리 파이 기반의 다중 카메라 시스템과 모듈형 프레임 구조를 채택하였다. 이 절에서는 시스템의 구성, 작동 원리, 데이터 흐름, 하드웨어 및 소프트웨어 통합 등 시스템 구조와 기능 전반을 다룬다.

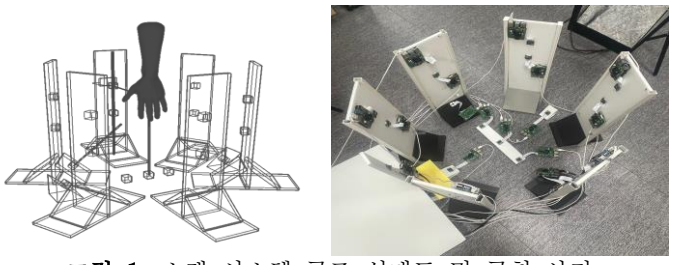


그림 1: 스캔 시스템 구조 설계도 및 구현 사진

시스템은 총 20대의 라즈베리 파이 4 Model B와 라즈베리 파이 카메라 모듈 3으로 구성되며, 각 장치는 병렬 이미지 캡처 및 전송을 수행한다. 각 라즈베리 파이에는 PoE(Power over Ethernet) + HAT이 장착되어 하나의 이더넷 케이블로 전원과 네트워크를 동시에 공급받으며, 모듈은 손의 다양한 각도를 촬영할 수 있도록 측면에 육각 기둥 형태로 배치되었다(그림 1).

3. 방법론

3.1. 제어 시스템 아키텍처 및 통신 구조

라즈베리파이 기반 다중 카메라 시스템의 제어를 위해,

본 연구에서는 메인 제어 컴퓨터(Main PC)와 각 라즈베리파이 장치를 PoE 네트워크 허브를 통해 연결하였다. 이 구성은 데이터 통신을 동시에 처리할 수 있어 배선의 복잡도를 낮추고, 안정적인 네트워크 환경을 구성하는 데 효과적이다. 네트워크 허브는 제어 PC와 직접 연결되어 전체 시스템을 하나의 로컬 네트워크로 통합한다. 각 라즈베리파이에는 고정 IP 주소가 설정되어 있어, 제어 PC에서는 해당 주소를 기반으로 각 장치에 정확히 접근할 수 있다. 제어 PC는 클라이언트로서 중앙에서 촬영 명령을 전송하고, 각 라즈베리파이는 서버 역할을 수행하여 명령 수신 후 카메라 촬영을 수행한다. 이후 촬영된 이미지는 라즈베리파이 내부에 임시 저장되거나, 필요 시 네트워크를 통해 중앙 서버로 전송된다.

3.2. 포토그래메트리 기법

본 시스템에 사용된 포토그래메트리 기법은 다수의 2D 이미지를 다양한 시점에서 촬영한 후, 이를 정합하여 3D 형상을 재구성하는 방식으로, 본 연구에서는 이를 활용하여 손의 3차원 모델을 생성한다. 이를 위해 시스템은 라즈베리 파이 기반의 다중 카메라 모듈을 육각 기둥 형태로 배치하여 다양한 각도에서 피사체(손)를 동시에 촬영할 수 있도록 설계되었다.
기존 선행 연구에서는 손 전체를 촬영하기 위한 카메라 구성은 이루어졌으나, 손가락 끝부분이나 손의 복잡한 굴곡 구조를 충분히 고려한 배치는 미흡하였다[2]. 이에 본 시스템은 손의 형태적 특징을 고려하여, 단순한 손바닥 중심의 평면 배치가 아닌 손가락 끝과 측면, 손등 등 복잡한 곡면 구조를 보다 정밀하게 촬영할 수 있도록 카메라를 측면뿐만 아니라 바닥에도 전략적으로 배치하였다.
또한, 포토그래메트리 알고리즘의 정합 정확도를 높이기 위해서는 모든 카메라가 동일한 시각에 이미지를 획득하는 것이 핵심이며, 이를 위해 본 시스템은 각 라즈베리 파이 장치를 NTP(Network Time Protocol) 기반으로 동기화하여 촬영 시각 간의 시간 오차를 최소화하고, 고정밀 다중 시점 데이터 수집이 가능하도록 구성하였다.

3.3. 스캔 데이터 획득 과정

사용자가 손을 프레임 중앙에 위치시킨 후 촬영 명령을 입력하면 제어 PC는 네트워크를 통해 각 라즈베리파이에 동시 촬영 명령을 전달하고, 각 장치는 카메라 모듈을 통해 이미지를 촬영한 뒤 해당 데이터를 저장하거나 서버로 전송한다. 이후 수집된 이미지는 포토그래메트리 정합 소프트웨어로 전달되어 3D 모델로 재구성된다.

4. 결과

본 연구에서 개발한 시스템은 총 20대의 라즈베리 파이 카메라를 활용하여 손을 다양한 각도에서 동시에 촬영하는 데 성공하였다(그림 2). 촬영된 이미지는 모두 네트워크를 통해 중앙 서버로 수집되었으며, 이후 Reality Capture 소프트웨어를 사용하여 3D 포토그래메트리 정합 과정을 수행하였다. 결과적으로 손 전체의 3D 메시 모델이 생성되었으며, 손가락 끝과 손등을 포함한 주요 구조의 형상은 비교적 잘 재현되었다.

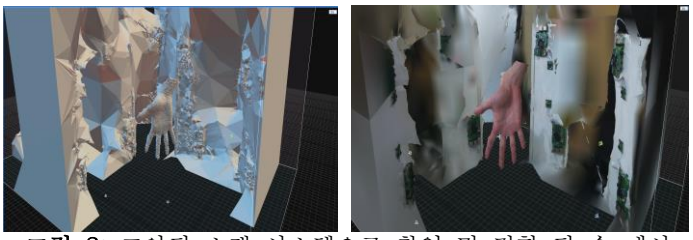


그림 2: 고안된 스캔 시스템으로 촬영 및 정합 된 손 메시

5. 한계점

본 논문은 저비용 손 스캔 시스템의 구현에 대한 예비 연구로써 그 가능성을 보여줄 수 있는 프로토타입을 제시하였다. 최신 관련 연구[2]와 비교하여 카메라의 전략적 배치를 통해 손 끝 부분에 대한 부분적인 개선이 있었지만 전체적인 복원 정확도가 50대의 카메라를 사용한 기존 연구의 성능에는 미치지 못했다. 일부 영역에서 촬영 시 광량 부족이나 최적화 문제 등으로 인해 노이즈가 발생하거나 메시가 불완전한 부분도 관찰되었다.

6. 결론

본 연구에서는 라즈베리 파이와 소형 카메라 모듈을 활용하여, 저비용·이동형 3D 손 스캐너 시스템을 설계 및 구현하였다. 제안된 시스템은 기존 선행 연구에서 사용된 수십 대 이상의 카메라 시스템과 비교하여 더 적은 수(20대)의 카메라만으로도 효과적인 손 형상 복원이 가능하도록 구성되었으며, 모듈형 프레임 구조를 채택함으로써 설치 유연성과 이동성을 확보하였다. 향후 연구에서는 시스템의 정합 정확도 향상을 위한 카메라 파라미터(초점 거리, ISO, 셔터 속도 등)의 최적화, 그리고 카메라 배치 구조의 개선을 통해 보다 정밀한 메시 복원 방법을 살펴볼 예정이다. 또한, 다양한 손 형태 및 조명 조건에서도 안정적인 스캔이 가능한 적응형 제어 알고리즘 및 실시간 피드백 기능을 추가하여, 실제 현장 적용 가능성을 높일 수 있는 방향으로 시스템을 고도화하고자 한다.

참고문헌

[1] Y. Yang, C. Kim, J. Jang, S. Park, and J. Lee, The development of a low-cost photogrammetry-based 3D hand scanner, *HardwareX*, Vol. 10, e00201, 2021.
[2] VynZ Research, *Global Handheld 3D Scanner Market Size by Product, Application, and Forecast*, Report ID: 438299, Published: August 2024

자율 주행 울타리를 이용한 군중 흐름 제어 시스템*

하영흠, 이다현, 아미레자, 김주란, 박채원, 최명걸
가톨릭대학교 미디어기술콘텐츠학과
heyongxin@catholic.ac.kr, {dahyun2608, tedtomson90, zerah.ox, dlfgnsdkd1004}@gmail.com, mgchoi@catholic.ac.kr

A Crowd Flow Control System Using Autonomous Mobile Fences

Yongxin He, Dahyun Lee, Amirreza, Juran Kim, Chaewon Park, Myung Geol Choi
Dept. of Media Technology & Media Contents,The Catholic University of Korea

요약

본 연구에서는 군중 밀집에 의한 위험을 방지하기한 동적 울타리 시스템을 제안한다. 유용성을 입증하기 위해 이태원 지형을 기반으로 한 H자형 경사 지형을 구축하고, 군중 밀도 임계값에 따라 자동으로 설치되고 위치를 변경하는 울타리 시스템을 적용하여 군중 밀집 상황 시뮬레이션 결과를 관찰하였다. 세 가지 조건 (울타리 미설치, 고정식 울타리, 밀도 기반 동적 울타리)을 비교한 결과, 인파 밀도가 특정 수준 이상으로 높은 경우 동적 울타리 시스템이 주요 통로의 혼잡을 효과적으로 줄이고 전체 대피 시간을 단축하는 데 가장 효과적인 것으로 나타났다.

1. 서론

도시 공공 공간에서 대규모 인파가 밀집하는 축제나 집회는 인명 사고의 위험을 높인다. 특히 좁은 통로에서 군중 밀도가 임계값을 초과하면 이동이 제한되고, 공간 압축 현상이 심해져 압사 사고로 이어질 수 있다. 2022년 이태원 압사 사고가 대표적인 사례이다. 기존의 군중 분산 및 입장 제한은 주로 인력이나 정적 차단 방식에 의존해, 실시간 대응이 어렵다는 한계가 있다. 이에 본 연구는 제한된 공간에서 자동으로 군중 흐름을 제어할 수 있는 자율 이동형 울타리 시스템을 제안한다. 자율주행 기술을 활용해, 평상시에는 비활성화된 위치에 대기하다가, 밀집도가 임계값을 초과하면 실시간 데이터에 따라 최적 경로로 이동하여 군중 흐름을 유도·차단하는 방식이다. 본 연구에서는 기존의 정적 울타리와 비교하여, 자율 이동형 울타리 시스템이 복잡한 도시 지형 및 급변하는 인파 상황에서 더욱 유연하고 효과적으로 군중을 제어할 수 있음을 시뮬레이션을 통해 정량적으로 평가하였다.

* 포스터 발표논문
* 본 연구는 경찰청 과학기술기반 군중밀집관리 기술 개발 연구 사업(RS-2024-00405100)의 지원으로 수행되었음

2. 방법 및 구현

2.1. 장면 및 기본 설계

시뮬레이션 환경은 이태원 사고현장과 유사하게 제작하였다. 해당 현장은 이태원로와 이태원27가길을 연결하는 길이 40m, 폭 4m 정도의 골목길이다[1]. 골목의 이태원27가길 방면은 이태원로 방면보다 지형 높이가 5m 더 높다. 전체 구조는 H자 모양이며 자세한 설계는 그림 1과 같다. 이태원로 (그림 1의 왼편)와 이태원27가길 (그림 1의 오른편)의 양 끝단을 군중 흐림이 시작되거나 끝나는 지점으로 설정하였다. 각 입구에서 다른 세 개의 출구로 이동이 가능하므로 전체 군중 흐름의 종류는 총 12가지이다. 본 연구에서는 성인, 어린이, 노인 세 가지 군중 유형을 설정하였고, 각 유형별로 신장, 어깨 너비, 이동 속도의 범위를 다르게 적용하였다.
이태원 사고 당시 사망자 명단에 나타난 연령 분포를 보면, 아동과 노인이 전체 인원의 약 8%를 차지하고 있었다. 이를 바탕으로 본 실험에서는 성인을 92%, 아동을 5%, 노인을 3% 비율로 설정해 시뮬레이션을 진행하였다. 군중 시뮬레이션 알고리즘은 RVO[2]를 사용하였다.

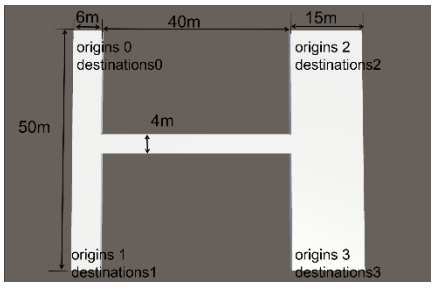


그림1: 이태원 사고현장을 모델로 제작한 H자형 지형 가상 환경

2.2. 울타리 제어 전략

우리가 제시하는 이동식 울타리의 우수성을 입증하기 위해, 본 연구에서는 정적 울타리를 사용하는 경우와 이동식 울타리를 사용하는 경우를 각각 실험하여 성능을 비교하였다. 정적 울타리는 행사 등으로 인파 집중이 예상되는 좁은 도로에서 흐름 통제를 위해 사전에 설치하

는 일반적인 방법으로, 주로 통로 입구에서 다소 떨어진 위치에 고정 설치된다. 이를 통해 보행자의 경로를 사전에 분산시키고, 일정한 순서로 대기하도록 유도하는 기능을 한다 (그림 2-왼쪽). 정적 울타리 경우 실시간으로 위치를 바꾸기 어렵기 때문에 흐름을 완전히 차단하는 배치는 할 수 없다.

반면, 이동식 울타리는 실시간 군중 밀도 정보를 기반으로 자유롭게 위치를 조정할 수 있어, 상황에 따라 일시적으로 흐름을 차단하거나 유도하는 보다 적극적인 제어가 가능하다. 따라서 통로 입구에 가까운 구간에서 배치 및 이동함으로써 다양한 경로에서 접근하는 보행자들이 순차적이고 안전하게 통과하도록 흐름을 능동적으로 조율하도록 하였다. 총 12개의 동적 울타리를 사용했으며 울타리의 크기는 두 가지이다. 짧은 것은 3개가 1열로 배치될 때 이태원27가길을 완전히 막을 수 있는 길이이며 긴 것은 3개가 1열로 배치될 때 이태원길을 완전히 차단할 수 있는 길이이다 (그림 2-오른쪽). 동적 울타리는 통로 주변 군중밀도가 4명/㎡에 도달하는 것이 감지되면 스스로 좁은 통로 가까이로 이동한다 (그림 4). 그 후 그림 3에서 설명된 세가지 개방 형태를 10초 간격으로 순차적 반복한다. 이후 군중 밀도가 2명/㎡ 미만으로 낮아지면 울타리 시스템은 자동으로 복귀한다.

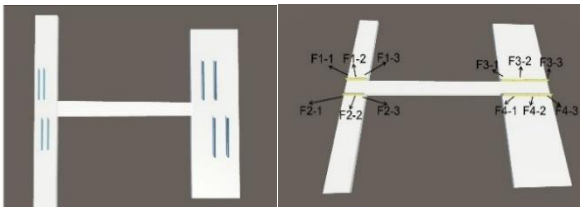


그림 2: (왼쪽) 실험에 사용된 정적 울타리 배치. (오른쪽) 실험에 사용된 동적 울타리가 모든 길을 차단하고 있는 상황

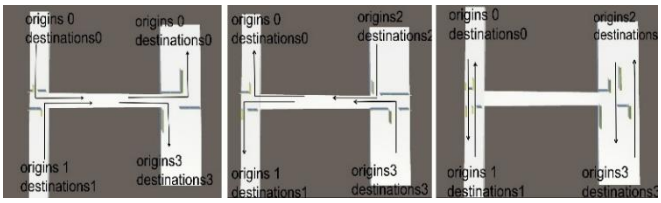


그림 3: 울타리 개방 방식 및 군중 이동 경로

3. 실험 결과 및 토의

3.1. 실험 결과 비교

그림 4: 높은 인구 밀도를 감지하고 스스로 최적의 설치 장소로 이동하여 울타리 실험 결과 (왼쪽에서 오른쪽으로 시간 순, 울타리 위치 검은 사각형으로 표시)

설정된 가성환경에서 인구 수를 다르게 하여 통로 부분의 인구가 모두 빠져나가는 데 걸리는 시간을 측정하였다. 저밀도 (1000~2000명)에서는 울타리 미설치 상황과동적 울타리 사용시의 대피 시간이 비슷했다. 정적 울타리는 항상 통로를 막고 있어 우회나 대기가 필요해 대피 시간이 가장 길고 효율이 가장 낮았다. 중밀도 (3000~6000명) 경우에는 울타리 미설치 경우에서 병목현상이 심해져 대피 시간이 크게 증가했다. 동적 울타리가 움직이며 군중 흐름을 다소 방해하여 약간 더 오래 걸렸다. 동적 울타리 적용시에는 약 1.0~1.4분의 추가 시간이 들었지만, 정적 울타리에 비해 대피 시간을 25~32% 단축시켰다. 고밀도 (7000명)에서는 울타리 미설치 경우의 대피 시간이 급격히 증가했다. 반면 동적 울타리 사용의 경우 5.5분으로 울타리 미설치 대비 약 24.6% 단축되었다. 인원이 7000명일 때 감지 구역의 밀도가 설정된 임계값을 초과하면, 동적 울타리가 작동하여 혼잡 구역을 일시적으로 차단하고 인파를 강제로 분산시킨다. 이를 통해 서로 다른 경로에서 오는 사람들 간의 충돌을 막고, 혼잡을 줄일 수 있다. 이러한 결과는 동적 울타리가 특히 인구 밀도가 높은 상황에서 효과적으로 작동한다는 점을 보여준다.

4. 결론

동적 울타리 시스템은 유연성과 효율성의 균형이 뛰어나며, 특히 고밀도 상황에서 우수한 군중 대피 성능을 보였다. 현재 실험에서는 동적 울타리의 이동 경로와 위치를 결정하기 위해 여러 차례의 시뮬레이션을 반복 관찰한 후, 사람의 직관에 기반하여 위치를 선정하였다. 따라서 본 연구에서 도출된 결과는 실험에 사용된 특정 환경과 인파 규모에 한정된 결과이다. 향후 연구에서는 새로운 지형과 실시간 군중 흐름을 반영하여 최적의 울타리 위치를 계산하여 동적으로 이동할 수 있는 자율적인 제어 알고리즘 개발을 목표로 한다.

표 1 인구 수에 따른 밀집 해소 시간 측정 결과

인원수 (명)	울타리 미설치 (min)	정적 울타리 (min)	동적 울타리 (min)
1000	1.30	2.16	1.40
2000	1.33	2.39	1.51
3000	1.38	3.26	2.44
4000	2.25	4.03	3.35
5000	3.08	5.28	4.49
6000	4.47	7.49	5.12
7000	7.29	8.17	5.50

참고문헌

[1] H. Liang, S. Lee, J. Sun, and S. Wong, "Unraveling the causes of the Seoul Halloween crowd-crush disaster," PLoS one, vol. 19, no. 7, p. e0306764, 2024.
[2] Van Den Berg, Jur, et al. "Reciprocal n-body collision avoidance." Robotics Research: The 14th International Symposium ISRR. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

포인트 클라우드에 대한 Heatmap 기반 3D 귀 랜드마크 탐지 GCN 모델

박평화 ° 이영성 한다성
한동대학교 휴먼테크융합과
(22473003, 22573001)@handong.ac.kr dshan@handong.edu

A Heatmap-Based GCN Model for 3D Ear Landmark Detection on Point Clouds

Pyeonghwa Park ° Youngsung Ree Daseong Han
Handong Global University

요약

본 연구는 곡률이 크고 개체 간 변이가 심한 3D 귀 포인트 클라우드에서 해부학적 랜드마크를 예측하기 위해 기존 그래프 합성곱 신경망(GCN) 기반의 히트맵 회귀 프레임워크를 도입한다. 기존 프레임워크가 얼굴 랜드마크에 기반한 것과는 달리 입력으로는 템플릿 정렬된 3D 귀 점군이 사용되며, Farthest Point Sampling(FPS) 및 k-NN 기반 그래프 구성을 통해 구조 정보를 학습한다. PAConv(Position-Adaptive Convolution)을 기반으로 한 GCN 구조를 통해 각 포인트의 특징을 추출하고[4, 5], 각 랜드마크에 대한 확률 분포(3D 히트맵)를 회귀한다. 회귀 된 히트맵은 MDS(Multidimensional Scaling) 및 soft-argmax 기반 후처리를 통해 정밀한 3D 좌표로 복원된다. 실험에서는 회귀 포인트 수와 히트맵 σ값을 조정하며 정량적 오차를 분석하였고, 그S 영향을 검증하였다.

1. 서론

3D 귀(ear)의 해부학적 랜드마크는 귀의 이륜, 이주, 귓불 등 고유한 구조를 나타내는 기준점들이며 정밀한 인식과 맞춤형 설계에 활용이 가능하다. 이를 정밀히 인식하는 기술은 가상 아바타 생성, AR 안경 및 VR 헤드마운트 기기의 사용자 맞춤 설계, 의료용 보청기 및 착용형 기기 디자인, 법의학적 신원 확인, 보안 생체인식 시스템 등 다양한 실제 응용 분야에서 핵심적으로 활용된다. 최근에는 GCN, Transformer 등 다양한 딥러닝 모델을 활용한 3D 얼굴(face) 랜드마크 감지 기술이 활발히 연구되고 있으나 [4], 이에 비해 귀 랜드마크 자동화 기술에 대한 연구는 매우 제한적인 수준에 머물러 있다 [1, 2]. 이는 귀가 얼굴에 비해 개인차가 크고, 비대칭 적이며 복잡한 곡률 구조를 가져 기존 얼굴 기반 모델을 그대로 적용하기 어렵기 때문이다. 이러한 특성으로 인해 현재까지는 전문가가 3D 스캔 데이터에 직접 랜드마크를 수기로 지정하는 방식에 의존하는 경우가 많다. 그러나 이 방식은 시간과 비용이 많이 소요될 뿐만 아니라, 작업자의 주관에 따라 비 일관성 문제를 야기하며 [2], 3D 귀 데이터를 활용한 기술의 확장성과 실용성을 저해하는 병목 요인으로 작용한다. 특히 Zhou et al. [3]에 따르면, 현재까지 공개된 귀 랜드마크 주석 데이터는 600여 개 수준에 불과하여 얼굴에 비해 학습 가능한 데이터의 수가 절대적으로 부족함을 보여준다. 이러한 데이터 희소성은 딥러닝 기반 귀 인식 및 랜드마크 추정 기술의 발전을 제한하는 중요한 요인 중 하나이다. 이에 본 연구에서는 이러한 한계를 극복하고, 정렬된 3D 귀 포인트 클라우드 상에서 기존에 잘 다뤄지지 않던 외이도(canal)까지 포함한 해부학적 랜드마크를 자동으로 추출하는 GCN 기반 히트맵 회귀 프레임워크를 제안한다.

2. 방법론

본 연구는 곡률이 크고 개체 간 변이가 심한 3D 귀 포인트 클라우드에서 해부학적 랜드마크를 예측하기 위해, 기존의 GCN 기반의 히트맵 회귀 프레임워크를 도입한다 [4]. 기존 프레임워크가 얼굴 랜드마크에 특화 되어있는 것과는 달리, 본 연구에서는 귀 랜드마크 탐지를 위해 총 296명의 20~30대 남녀 한국인 및 북미인으로 구성된 포항공과대학교 자체 수집 3D 귀 데이터를 사용하였다. 각 샘플에서 FPS(Farthest Point Sampling)를 통해 2048개의 대표 점을 추출한 후, k-NN을 기반으로 그래프를 구성한다. 이후 weight bank 8개를 사용하는 PAConv 4계층을 통해 각 점의 특징 벡터(채널 수 64)를 추출하고 이를 1024채널로 확장한 뒤 전역 정보를 결합하였다. 최종적으로 후처리 MLP를 통해 채널을 랜드마크 개수만큼 축소하여 회귀를 수행하였으며, 모델은 학습률 0.003, 배치 크기 8, 총 250 epoch의 설정으로 학습하였다. 또한 Adaptive Wing Loss를 사용해 히트맵 회귀 시 작은 오차에도 민감하게 학습하도록 한다. 회귀 된 히트맵은 soft-

* 포스터 발표논문

argmax와 MDS 기반 후처리를 통해 각 랜드마크의 정밀한 3D 좌표로 복원된다 [4]. 본 연구에서는 프레임워크의 하이퍼파라미터인 회귀 포인트 수와 히트맵 σ값이 예측 정확도에 미치는 영향을 살펴본다.

3. 실험 및 결과

기존 귀 데이터 296개 중 정렬화 된 284개 샘플을 기반으로 수행되었으며, 각 샘플은 40개의 랜드마크로 구성되어 있다. 실험은 히트맵의 σ값과 회귀에 사용되는 포인트 수(regression point)를 달리하며 총 4가지 조건에서 수행되었다.

	10 Regression Points	40 Regression Points
σ=5	3.8514	5.0685
σ=10	5.3887	5.2320

표 1: σ 값과 회귀에 사용되는 포인트 수에 따른 ME 값

Model	Landmark Number	Mean Error (mm)
Ours	40(with Canal)	3.8514
O'Sullivan et al. (2020)	13(without Canal)	2.086

표 2: 기존 연구와의 정량적 성능 비교 (ME mm)

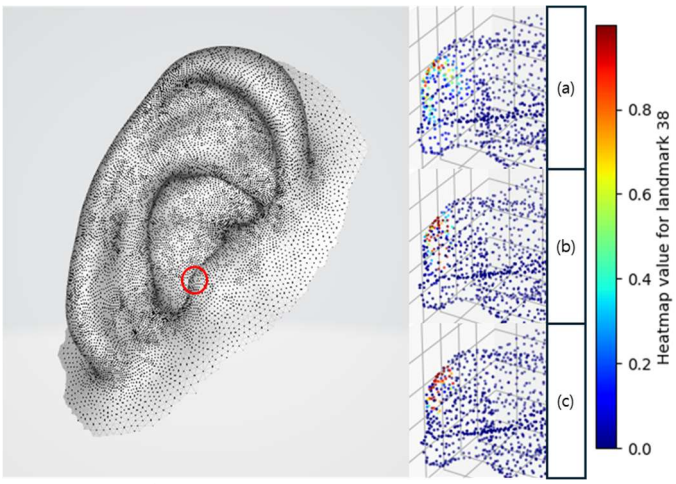


그림 1: Tragus landmark point heatmap
(a) GT origin dataset,
(b) σ = 5 & 10 Regression points,
(c) σ = 5 & 40 Regression points

[표 2]에서 볼 수 있듯 본 연구의 최적 설정(ME=3.8514 mm)은 기존 귀 랜드마크 연구인 PointNet++ 기반 O'Sullivan et al. [2]의 평균 오차(2.086 mm) 대비 다소 높은 수치를 보였으나 이는 외이도(canal)를 포함한 더 복잡한 해부학적 구조까지 예측 대상에 포함되었기 때문이다. 실제로 O'Sullivan et al. [2]은 총 13개의 랜드마크 만을 대상으로 하며, 외이도 내부 구조 랜드마크는 제외하고 있어 본 연구와는 예측 난SSSS이도 및 적용 범위에서 차이가 존재한다. 위 결과는 [표 1]에 정리되어 있으며 σ=5, 회귀 포인트 수가 10일 때 가장 낮은 평균 오차(ME)를 나타낸다. 이는 작은 크기의 정밀 패치를 중심으로 회귀를 수행하는 것이 예측 정확도에 긍정적인 영향을 미칠 수 있음을 시사한다. 반면 σ가 커질수록 히트맵이 퍼지며 예측 중심성이 낮아져 오차가 증가하는 경향을 보였다. [그림 1]에서는 Tragus 랜드마크에 대해 GT 데이터셋과 두 가지 실험 설정 간의 히트맵 결과를 시각화 하였다. (a)는 GT 원본, (b)는 σ=5, 회귀 포인트 10개 조건, (c)는 σ=5, 회귀 포인트 40개 조건이며, 히트맵의 응답 강도 분포가 패치 크기에 따라 달라지는 양상을 확인할 수 있다. 본 연구는 회귀 포인트 수와 히트맵 σ 값이 3D 귀 랜드마크 예측 정확도에 미치는 영향을 실험적으로 정량 분석하였다. 이를 통해 정밀한 예측을 위한 패치 크기와 회귀 포인트 수의 조합을 제시하며 향후 고정도 귀 분석 모델의 파라미터 설정에 유용한 기준을 제공한다.

참고문헌

[1] J. Lei, X. You, and M. Abdel-Mottaleb, "Automatic ear landmark localization, segmentation, and pose classification in range images," IEEE Trans. Syst. Man Cybern. Syst., vol. 46, no. 2, pp. 165-177, 2016.
[2] E. O'Sullivan and S. Zafeiriou, "3D landmark localization in point clouds for the human ear," in Proc. FG, pp. 402-406, 2020.
[3] Y. Zhou and S. Zafeiriou, "Deformable models of ears in-the-wild for alignment and recognition," in Proc. FG, 2020.
[4] Y. Wang et al., "Learning to detect 3D facial landmarks via heatmap regression with graph convolutional network," in Proc. AAAI, vol. 36, no. 2, pp. 2595-2603, 2022.
[5] M. Xu et al., "PAConv: Position Adaptive Convolution With Dynamic Kernel Assembling on Point Clouds," in Proc. CVPR, pp. 3173-3182, 2021.

한국컴퓨터그래픽스학회 2025 학술대회 논문집

발행인 최수미

편집인 장 윤

발행처 사단법인 한국컴퓨터그래픽스학회

주소 05006 서울특별시 광진구 능동로 209 (군자동)
세종대학교 광개토관 1011호

전화 02-6935-2539

URL http://www.cg-korea.org/

KCGS2025

GENERATING WORLDS, RENDERING REALITY

Platinum Sponsor

CLO Virtual Fashion

Gold Sponsor

OSSTEM IMPLANT

Silver Sponsors

IITP 정보통신기획평가원

세종대학교 초실감 XR 연구센터 EXTENDED REALITY RESEARCH CENTER

SW중심대학사업단 HALLYM UNIVERSITY

BARO Bring AI Right ON

가톨릭대학교 컴퓨터그래픽스연구실 COMPUTER GRAPHICS LAB

DIBLAT

KAIST Visual Media Lab.

세종대학교 SW중심대학사업단

Bronze Sponsors

GENGENAI

고려대학교 ICT명품인재양성사업단 Korea University ICT Creative Consilience Foundation

Simulated Reality Ewha ITRC

숭실대학교 Soongsil University

메타버스융합대학원 Metaverse Convergence Graduate School