일반논문 (Regular Paper)
방송공학회논문지 제29권 제6호, 2024년 11월 (JBE Vol.29, No.6, November 2024)
https://doi.org/10.5909/JBE.2024.29.6.1010
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

버추얼 프로덕션을 위한 딥러닝 기반 2.5D 에셋 생성 기술 연구

추 해 수[a], 진 인 환[a], 정 성 훈[b], 김 정 환[a)‡], 공 경 보[b)‡]

Deep Learning-Based 2.5D Asset Generation Techniques for Virtual Production

Haesoo Choo[a], In-hwan Jin[a], Seong-Hun Jeong[b], Junghwan Kim[a)‡], and Kyeongbo Kong[b)‡]

요 약

본 논문은 기존의 3D 복원 알고리즘을 검토하여 이들의 한계를 확인하고, 2D 이미지 생성 및 편집 모델을 활용하여 3D 공간을 효율적으로 표현할 수 있는 개선된 2.5D 데이터 생성 알고리즘을 제안한다. 2.5D 데이터는 2D 이미지로부터 카메라 각도에 따라 변화하는 장면과 사물의 움직임을 현실적으로 제공하며, 물과 같은 유체의 애니메이션을 포함한 배경 에셋을 생성하여 몰입감을 높이고자 한다. 이 과정은 총 5단계로 구성되며, 기존 연구들을 비교하여 최적의 결과를 도출하는 방법을 제시한다. 각 단계에서는 지속적으로 발전하고 있는 기존 연구들을 추가하여 제안하는 알고리즘의 확장 가능성을 보여준다. 이를 통해 AI를 활용하여 낮은 비용과 적은 시간 내에 높은 퀄리티의 에셋 데이터를 생성할 수 있는 접근을 제안한다. 이 기술은 실제 영화 촬영 산업에서 중요한 역할을 할 것으로 기대되며, 부족했던 에셋 데이터셋 공급을 충족시키는 중요한 기술적 해결책으로 자리 잡을 것이다.

Abstract

This paper proposes an improved 2.5D data generation algorithm that efficiently represents 3D spaces using existing 2D image generation and editing models. 2.5D data realistically provides scenes and object movements that vary with camera angles from 2D images, aiming to enhance immersion through generating background assets including fluid-like animations. The process consists of five stages, where existing research is compared to derive optimal results at each step. By incorporating ongoing advancements in each proposed stage, the algorithm demonstrates potential for expansion. This approach suggests using AI to generate high-quality asset data efficiently and affordably, and it is expected to play a crucial role in the film production industry by addressing shortages in asset datasets.

Keyword : Virtual Production, Deep Learning, 2.5D Asset Generation, In-camera VFX, Parallax Effect

a) 부경대학교 미디어커뮤니케이션학부 휴먼ICT융합전공(Division of Media School, Pukyong National University)
b) 부산대학교 전기전자공학부 전자공학전공(Department of Electrical & Electronics Engineering, Pusan National University)
‡ Corresponding Author : 공경보(Kyeongbo Kong), 김정환(Junghwan Kim)
E-mail: kbkong@pusan.ac.kr, media.jhk@gmail.com
Tel: +82-51-510-2399, +82-51-629-5479
ORCID: https://orcid.org/0000-0002-1135-7502, https://orcid.org/0000-0001-5360-0059

※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00456152) and a New Faculty Research Grant of Pusan National University, 2023 and regional broadcasting development support project funded by the Foundation for Broadcast Culture.
· Manuscript September 6, 2024; Revised October 16, 2024; Accepted October 16, 2024.

Ⅰ. 서 론

배우가 녹색 천 앞에서 연기한 후 컴퓨터 그래픽(CG) 배경을 합성하는 '그린 스크린' 제작 방식은 후처리에 많은 시간과 비용이 들기 때문에 영상 콘텐츠 제작에서 고비용 투자가 필요하다. 예를 들어, 거의 모든 장면에서 CG가 사용된 영화 <아바타2>에는 2000명의 시각 효과 기술자가 3년 동안 투입되었고, 초당 2.3억 원의 제작비가 사용되었다. 이러한 고비용의 그린 스크린 방식의 대안으로, 사전에 만든 배경(에셋)을 상황에 맞게 LED Wall에 출력해 그 앞에서 연기 및 촬영하는 방식인 버추얼 프로덕션이 최근에 등장하였다. 이 방식은 후처리 작업을 크게 줄여 비용을 절감하고, 배우들이 주변 상황에 몰입해 연기할 수 있어 기존 콘텐츠 제작 산업은 물론, 최근 떠오르는 OTT 콘텐츠 제작 산업에서도 매우 각광받는 촬영 방식이다.

버추얼 프로덕션에서는 현실 장면을 3D 가상 공간에 구현해야 하므로, 상황에 적합한 3D 에셋 데이터셋 구축이 중요하다. 2019년에 등장한 인카메라 VFX(In-camera VFX)는 버추얼 프로덕션의 핵심 기술로, 언리얼 엔진을 이용해 사전 제작한 3D 에셋을 LED Wall에 실시간으로 렌더링한다. 이 기술은 시차 효과를 반영해 가까운 물체는 빠르게, 먼 물체는 천천히 움직여 현실감을 제공한다. 그러나 기존의 게임 엔진을 사용한 3D 에셋 데이터셋 제작 방식은 시간과 비용이 많이 소요되어 에셋 데이터셋 공급 부족 문제가 발생되고 있으며, 모든 3D 에셋을 LED Wall에 실시간으로 렌더링하기 위해 고사양 장비가 필요하다.

최근에는 딥러닝과 신경망 기술을 기반으로 생성적 AI(Generative AI)가 활발히 연구되고 있다. 생성적 적대 신경망(GANs, Generative Adversarial Networks)[1]과 같은 기술은 이미지와 영상을 효율적으로 생성하고 변환하는 데 기여하고 있다. 이러한 AI 기술은 전통적인 방법보다 비용과 시간 소모를 크게 줄여 산업에서도 널리 사용되고 있다. 최근 연구에서는 기존의 시간과 비용이 소요되는 3D 에셋 데이터 제작 방식을 대신하여 보다 효율적으로 에셋을 제작하기 위해 다양한 연구가 진행되고 있다.

본 논문에서는 AI를 활용한 3D 에셋 데이터 제작 기술 동향을 조사하고, 각 기술의 한계를 분석한 후, 2D 이미지 생성 및 편집 모델을 활용한 효율적인 3D 에셋 데이터 제작 알고리즘을 제안하고자 한다.

Ⅱ. AI를 활용한 3D 공간 복원 및 표현 기술 동향

1. AI 기반의 3D 공간 복원 방법

3D 공간을 복원하는 것은 그림 1에서 보여주듯이 다양한 방법을 통해 연구되고 있다.

1.1 파노라마 이미지로부터 3D 공간 복원 기술

HorizonNet[2]은 단일 파노라마 이미지로부터 3D 공간의 레이아웃을 추정한다. 이 방법은 이미지에서 실내 공간의 모서리를 추정하고, 교차하는 벽이 서로 수직인 맨해튼 세계로 가정하여 3D 공간을 복원한다. 그러나 의자나 소파와 같은 물체는 무시되고 벽만 복원된다는 한계가 있다. 이를 해결하기 위해 PanoContext-Former[3]와 3D Room[4]과 같은 후속 연구들이 등장했다. 이들은 파노라마 이미지로부터 공간의 레이아웃을 생성하고 물체를 포함한 3D 공간 복

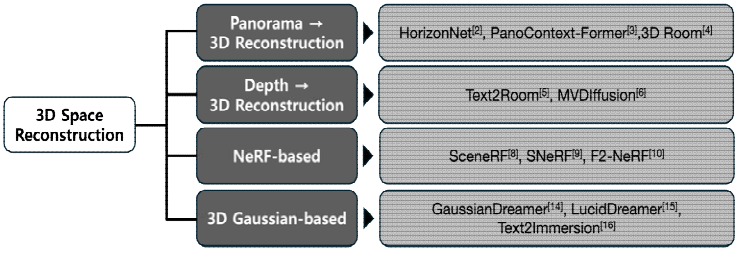

그림 1. AI 기반 3D 공간 복원 방법 구분
Fig. 1. AI-based 3D space restoration methods

원을 시도했지만, 맨해튼 세계나 평행한 공간과 같은 특정 조건의 실내 환경에만 적용될 수 있는 제한이 있다.

1.2 깊이 정보를 통한 3D 공간 복원 기술

Text2Room[5]과 MVDiffusion[6]은 깊이 정보를 활용하여 다양한 장면의 3D 공간을 복원하는 연구를 진행하였다. Text2Room은 입력된 텍스트를 기반으로 이미지를 생성하고, 깊이 정보를 이용해 3D 공간을 복원했다. 그러나 이 과정에서 물체가 중복 생성되거나 잘못된 깊이로 인해 결과물이 불완전해지는 문제가 있었다. MVDiffusion은 이러한 문제를 개선하기 위해 파노라마 이미지를 생성하고 깊이 정보를 샘플링하여 기존 연구를 발전시키려 했다. 그러나 이들 접근법은 깊이 정보가 한정된 실내 공간에만 유효하며, 실외와 같이 깊이 정보가 무한한 공간에는 적합하지 않을 수 있다.

1.3 NeRF를 기반으로 한 3D 공간 복원 기술

2020년에 등장한 Neural Radiance Fields(NeRF)[7]는 신경망을 이용해 고해상도의 3D 장면을 재구성하고 새로운 시점에서 이미지를 생성하는 기술이다. NeRF는 각 픽셀을 통과하는 광선(ray)을 카메라의 방향과 위치를 기반으로 정의하여 방사율(radiance)과 밀도(density)를 계산하고, 이를 통해 3D 장면을 재구성한다. SceneRF[8], Snerf[9], F2-nerf[10] 등은 NeRF를 기반으로 한 모델로, 3D 공간 복원에 중점을 두고 연구되었다. SceneRF는 구면 투영이 평면에 비해 왜곡이 적은 특성을 활용하여 카메라 시야를 확장하고, 입력 이미지의 시야 밖 색상과 깊이를 예측하여 실내의 3D 공간을 복원한다. F2-nerf는 임의의 카메라 궤도를 입력으로 받아 실외와 같은 넓은 범위의 3D 공간을 자연스럽게 표현하는 방법을 연구한다. 그러나 NeRF 기반 모델들은 고해상도 이미지 생성을 위해 긴 훈련 시간이 필요하고, 실시간 렌더링이 어렵다. 또한 복잡한 장면 표현에 한계가 있어 다양한 장면과 현실적인 몰입감을 요구하는 배경 에셋으로 활용하기에는 제한이 된다.

1.4 3D Gaussian Splatting 기반 3D 공간 복원 기술

가장 최근 기술 중 하나인 3D Gaussian Splatting[11]은 3D 공간에서 여러 개의 3D Gaussian이 모여 하나의 장면을 구성하는 기술이다. 이 기술은 NeRF와 유사하게 여러 이미지와 촬영 위치 정보를 활용하여 3D 장면을 재구성한다. 3D Gaussian은 명시적 표현 방법으로, 고해상도 기준의 State Of The Art(SOTA)인 Mip-NeRF[12]보다 우수하며, 학습 시간 기준 SOTA인 InstantNPG[13]보다 짧은 학습 시간을 제공할 수 있다. 이 모델은 적은 계산량으로도 높은 품질의 재구성 결과를 보장하며, 복원된 3D 공간을 2D 이미지로 효율적으로 프로젝션할 수 있다. 이러한 3D Gaussian을 기반으로 한 GaussianDreamer[14], LucidDreamer[15], Text2Immersion[16] 등의 기술이 텍스트 기반 3D 공간 복원을 시도하고 있다. GaussianDreamer는 기존의 diffusion 모델을 활용하여 객체뿐만 아니라 배경까지 생성할 수 있으며, 또한 Lucid-Dreamer는 다양한 카메라 시점에서 3D 공간을 복원한다. Text2Immersion은 몰입감 있는 3D 공간을 위해 깊이 있는 카메라 궤적을 제공한다. 그러나 3D Gaussians을 통한 공간 복원은 상용화에는 한계가 있으며, 현실적인 디테일에서 아티팩트가 발생할 수 있다.

2. 2.5D를 통한 3D 공간 표현 방법(Cuebric)

완전한 3D 공간을 복원하는 기술에는 여전히 해결해야 할 기술적 문제가 있다. 실제 상용화를 위해서는 학습 시간과 에셋 품질을 개선할 필요가 있다. 반면, 2.5D 데이터를 생성하여 3D 공간을 표현하는 방법은 기존 2D 이미지 모델을 활용해 높은 품질을 보장할 수 있는 효율적인 접근 방식이다. 2.5D 데이터는 단일 이미지 내의 레이어를 분리하여 다양한 카메라 시점에서 시차 효과를 통해 입체감과 깊이감을 표현한다. 이는 3D 공간을 현실적으로 보여주면서도 2D 이미지를 통해 효율적으로 표현할 수 있는 방법이다.

관련된 선행 연구로 Seyhan Lee가 있다. 2020년에 설립된 이 회사는 크리에이티브 AI 프로덕션에 주력하며, 영화 및 엔터테인먼트 산업에서 AI를 활용한 다양한 프로세스를 개발하고 있다. Seyhan Lee는 특히 버추얼 프로덕션을 위한 AI 도구인 Cuebric[17]을 개발하여 2.5D 데이터 생성을 통해 효율적인 배경 에셋을 제공하고 있다. 그림 2에서 확인할 수 있듯이, Cuebric은 2D 이미지를 기반으로 3D 공간을 효과적으로 표현하기 위해 다섯 가지 주요 과정을 거친다.

먼저 Image Generation 단계에서 2D 이미지를 생성한다.

그림 2. Seyhan Lee에서 제공되는 Cuebric의 프레임워크
Fig. 2. Cuebric's framework from Seyhan Lee

그 후, Image Segmentation 단계에서 이미지를 배경과 전경 등 여러 레이어로 분할한다. 분할된 이미지에는 Inpainting을 수행하여 가려진 영역을 채우고 새로운 장면을 생성하는 Image Inpainting 단계가 이어진다. 이후 Inpainting이 완료된 이미지에 Outpainting을 적용하여 각각 넓은 배경과 전경을 포함하는 개별 레이어를 생성한다. 마지막으로, 분리된 이미지들을 3D 콘텐츠 편집 프로그램에 불러와 깊이 정보를 기반으로 2.5D 데이터를 사용해 3D 공간을 표현한다.

2.1 Image Generation

Cuebric에서는 입력 프롬프트를 통해 사용자가 원하는 이미지를 생성할 수 있으며, negative prompt나 기타 파라미터를 조절해 사용자의 제어 가능성을 향상시켰다. 또한 생성 모델 서비스들의 베이스 모델로 주로 사용되는 Stable Diffusion[18] 모델을 추가 학습하여 다음 4가지의 이미지 생성 모델을 제공한다.

- Standard model: 사용자가 다양한 파라미터를 조절해 고품질의 영화 이미지를 생성한다.
- Classic model: 부드러운 안개와 광채 등 따뜻한 색상의 밝은 이미지를 생성하는 데 특화된 모델이다.
- Moody model: 자연스러운 조명과 낮은 노출을 활용하여 거친 고대비 이미지를 생성한다.
- Sci-fi model: 최첨단 LED 조명과 정밀한 카메라 데이터를 통해 선명한 이미지를 생성한다.

이미지 퀄리티에 따라 생성 시간은 약 10초 정도 걸리며, 사용자가 작성하는 프롬프트와 negative prompt에 따라 다양한 결과를 얻을 수 있다.

2.2 Image Segmentation

입력된 단일 이미지에서 배경과 전경을 여러 레이어로 분리하기 위해 Cuebric에서는 이미지 분할 기능을 제공한다. 이 기능은 원본 이미지나 Cuebric의 이미지 생성 모델을 통해 생성한 이미지에 적용할 수 있다. 구체적으로 두 가지 이미지 분할 기능이 있다.

- Semantic Segmentation: 이미지 내의 모든 객체를 픽셀 단위로 분리한다. 이 기능은 다음 두 가지 하위 기능으로 사용자 편의성을 높인다.
- Hover and Click: 마우스로 클릭하여 원하는 객체를 직접 분할할 수 있다.
- Segment All Object: 이미지 안의 모든 객체를 한 번에 분할할 수 있다.
- Depth Segmentation: 이미지에서 깊이 정보를 추출하고, 특정 깊이 범위를 선택하여 원하는 깊이의 객체만 남길 수 있다.

2.3 Image Inpainting

이미지 분할 이후 가려져 있던 영역에 대한 비어 있는 부분이 드러난다. 이러한 레이어를 배경 에셋으로 사용할 경우, 입력 이미지와 동일한 시점에서는 완전한 배경이 보

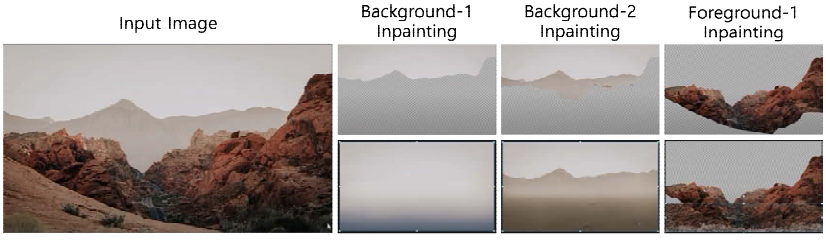

그림 3. Cuebric에서 제공하는 Image Inpainting 결과
Fig. 3. Image Inpainting results provided by Cuebric

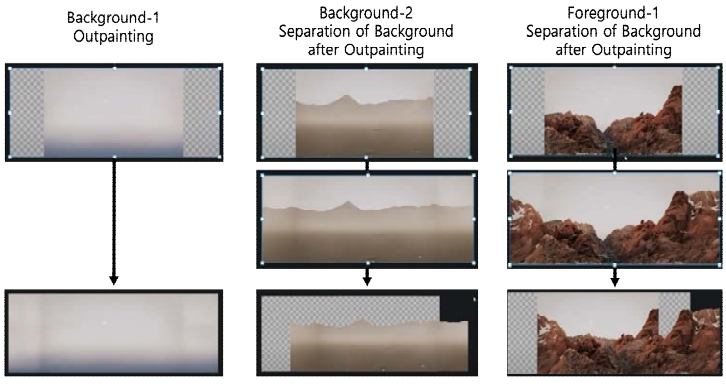

그림 4. Cuebricr 기반의 Image Outpainting 결과
Fig. 4. Image Outpainting results with Cuebric

이지만, 카메라 시점이 조금만 변경되면 비워진 부분이 노출되어 현실감이 떨어진다. 따라서 다양한 카메라 시점에서 자연스럽게 보이기 위해서는 분리된 레이어에 대한 Inpainting 기술이 필수적이다. Cuebric은 가려진 영역에 대해 Inpainting을 수행하며, 기본 툴을 사용해 원본 이미지의 영역을 지우고 텍스트로 새로운 객체를 생성하는 기능도 제공한다. 그림 3은 입력 이미지에서 깊이 정보를 기반으로 분리된 개별 레이어의 Inpainting 결과를 보여준다.

전경 1의 경우, 전경 1보다 앞선 영역에 대해서만 Inpainting을 진행된다. 이는 가려진 영역에 대해서만 Inpainting을 수행하기 때문이다.

2.4 Image Outpainting

생성된 이미지는 가로 비율이 비교적 짧아 실제 배경 에셋으로 사용하기에는 부족하다. 이를 해결하기 위해 Cuebric에서는 확장하고자 하는 영역을 설정하고, 텍스트를 입력하여 원본 이미지와 자연스럽게 이어지는 이미지를 추가로 생성하는 Outpainting 서비스를 제공한다. 그림 4는 Cuebric 기반의 Image Outpainting 결과를 보여준다. Outpainting을 진행하기 위해서는 분리된 레이어에 임의의 배경을 설정한 뒤, Outpainting 수행하고 다시 배경과 사용하고자 하는 레이어를 분할하여 최종 개별 레이어를 생성한다.

2.5 2.5D 구성

그림 5는 2.5D 데이터를 생성하기 위한 마지막 단계로, 생성된 최종 개별 레이어를 언리얼 엔진에 가져와 2.5D를 구성한다. 이때 사용자가 설정한 깊이에 따라 레이어를 배치하여 보다 현실적인 몰입감을 제공한다. 이를 통해 사용자는 2D 이미지로 2.5D 데이터를 생성할 수 있으며, 2.5D 데이터는 2D 이미지만으로도 3D 공간의 시차 효과와 다양한 몰입감을 제공한다. 이는 기존의 3D 공간 복원 방법보다 효율적이며 훨씬 빠른 생성 시간으로 3D 공간을 표현할 수 있게 한다.

그림 5. 언리얼 엔진 기반의 2.5D 구현 결과
Fig. 5. 2.5D result with Unreal Engine

Ⅲ. AI 기반 배경 에셋 생성 기술

Cuebric은 2.5D 데이터를 이용해 효율적으로 3D 공간을 표현하고 배경 에셋을 생성한다. 그러나 사용자 제어 도구는 입력 프롬프트와 마우스 클릭에 제한되어 있어 세밀한 제어가 어렵다. 또한 각 단계별 생성 모델이 공개되지 않아 사용자가 자체 데이터셋으로 재학습할 수 없어 상용화에 어려움이 있다.

최근 AI를 이용한 2D 이미지 생성 및 편집 기술이 급격히 발전하고 있으며, 이는 개선된 결과를 이끌어내고 있다.

다양한 모델들이 2D 이미지 생성 및 편집 기술을 기반으로 연구되고 있어, 더 나은 성능을 기대할 수 있다. 우리는 상용화 가능한 배경 에셋 데이터를 효율적으로 생성하기 위해 이러한 연구를 진행하고 있으며, 기존 Cuebric의 한계를 극복하기 위해 2D 이미지 생성 및 편집 모델을 활용한 알고리즘을 제안한다.

1. 제안하는 프레임워크

그림 6은 제안하는 2.5D 데이터 생성 알고리즘의 전체적인 프레임워크를 보여준다. 첫 번째 단계인 Image Generation에서는 다양한 조건과 입력 프롬프트를 활용해 사용자가 원하는 2D 이미지를 생성한다. 이후 프레임워크에서는 Cuebric과 달리 Outpainting을 선행하여 생성된 이미지를 배경 에셋으로 사용할 수 있도록 비율을 조정한다. 다음으로, Image Segmentation을 통해 이미지 내의 레이어를 분리하고, 이를 통해 여러 개의 배경과 전경 레이어를 얻는다. 각 레이어에 대해 다양한 카메라 각도에서 보이지 않았던 장면을 나타내기 위해 Image Inpainting을 수행하여 현실적인 몰입감을 제공한다. 마지막으로 각 레이어를 깊이 정보를 기반으로 3D 공간을 표현한다.

대부분의 배경 에셋은 물과 구름 등 움직임이 있는 풍경으로, 이러한 유체가 움직이지 않을 때 사용자들이 느끼는 몰입감이 떨어질 수 있다. 이를 해결하기 위해 기존의 애니메이션 생성 모델을 활용하여 움직임이 있는 배경 에셋을 생성한다. 생성된 이미지에 움직임을 적용해 애니메이션 비디오를 생성함으로써 더욱 다양한 배경 에셋을 사용자에게 제공할 수 있다.

우리가 제안하는 프레임의 각 단계에서 기존의 기술을 비교하여 최선의 결과를 얻고자 한다. 그림 7은 각 단계별로 사용한 알고리즘을 정리한 것으로, 알고리즘에 대한 상세 기술과 실험 결과를 비교하였다.

2. 단계별 상세 기술

2.1 Image Generation

우리는 사용자가 원하는 배경 에셋을 만들기 위해 다양한 이미지 생성 모델을 활용한다. 먼저 Diffusion 모델 기반의 Stable Diffusion을 사용하여 입력 프롬프트로 고해상도 이미지를 생성한다. 이 모델은 노이즈를 점진적으로 추가하고 제거하는 방식으로 이미지를 생성하며, 오픈 소스로 공개되어 다양한 연구자와 개발자에게 널리 사용되고 있다. 우리는 Stable Diffusion 2.1과 최근 출시된 Stable Diffusion XL을 사용한다. Stable Diffusion XL은 수백억 개의

그림 6. 본 연구에서 제안하는 2.5D 에셋 생성 프레임워크
Fig. 6. 2.5D asset creation framework proposed in this study

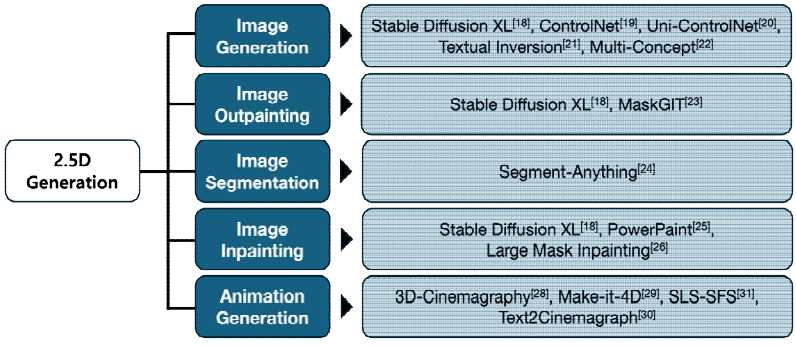

그림 7. 2.5D 데이터 생성을 위해 사용한 단계별 알고리즘
Fig. 7. Steps of the algorithm used for generating 2.5D data

그림 8. 동일한 입력 프롬프트에 대한 Stable Diffusion 2.1 버전과 Stable Diffusion XL 버전 비교 결과
Fig. 8. Comparison results of Stable Diffusion 2.1 version and Stable Diffusion XL version for the same input prompt

매개변수를 가지고 있어 더 복잡한 패턴과 세부 사항을 학습하여 높은 퀄리티의 이미지 생성을 가능하게 한다. 특히 복잡하고 디테일한 이미지 생성에서 우수한 성능을 발휘하는 것을 그림 8을 통해 알 수 있다.

우리는 사용자가 원하는 이미지를 생성하기 위해 입력 프롬프트 외에도 깊이, 스케치, 앳지 등의 추가적인 조건을 입력으로 받는 ControlNet[19]을 사용한다. 그림 9는 제공한 스케치 정보와 입력 프롬프트를 통해 생성한 이미지 결과이다. ControlNet은 각 조건에 대한 인코더를 개별적으로 학습하여 각 조건을 독립적으로 처리한다. 이전에는 하나의 조건과 프롬프트로 이미지를 생성했지만, Uni-controlNet[20]을 도입함으로써 여러 조건을 복합적으로 사용하여 더 구체적인 이미지를 생성할 수 있게 되었다.

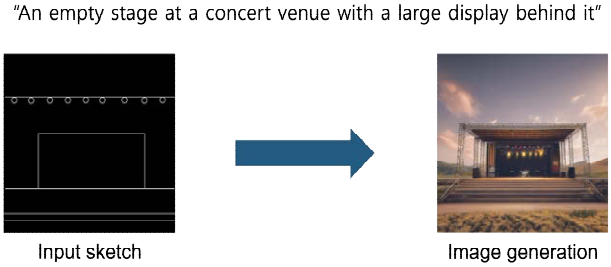

그림 9. 사용자가 제공하는 스케치와 입력 프롬프트를 함께 사용하여 이미지를 생성한 ControlNet 결과
Fig. 9. ControlNet results generated using the user's provided sketch and input prompt

장면에 따라 이전 배경 에셋과 현재 배경 에셋 간의 일관성이 필요할 수 있다. 기존 이미지 생성 모델은 동일한 입력 프롬프트나 조건으로 매번 다른 이미지를 생성하기 때문에 이러한 요구를 충족하기 어렵다. 이를 해결하기 위해 textual 정보를 담은 단어 S*을 학습시켜 비슷한 분위기의 이미지를 생성하는 Textual Inversion[21] 기술이 연구되었다. 이 기술은 Text에서 영상을 만들어내는 과정의 역과정으로, 이미지에서 Text를 뽑아내는 Inversion 기술을 사용한다. 그림 10에서 보여지듯이 이전 에셋에서 정보를 추출하여 일관성 있는 다른 에셋을 생성할 수 있다. 최근 Multi-Concept는 두 개의 S*을 혼합하여 이전 에셋의 정보를 기반으로 더 일관성 있는 이미지를 생성하는 기술을 제안하였다.

그림 10. 입력 이미지로부터 Textual 정보를 담은 단어로 생성한 Texture Inversion 결과
Fig. 10. Texture Inversion results generated from input images into words containing textual information

우리는 기존의 다양한 2D 이미지 생성 모델을 활용해 사용자가 원하는 이미지를 보다 효과적으로 생성하고자 한다.

2.2 Image Outpainting

우리는 각 단계에서 "Off-the-shelf-model"을 사용하여 특별한 맞춤화 없이 즉시 적용할 수 있도록 한다. 이로 인해 기존의 고퀄리티 모델뿐만 아니라 최근 연구된 다양한 모델도 활용할 수 있다. 일반적으로 이미지 생성 모델은

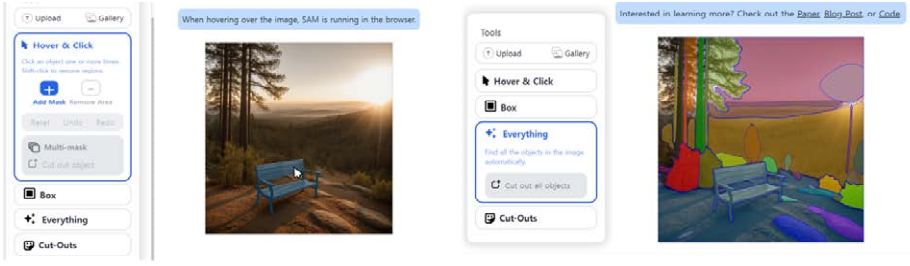

그림 11. Segment-Anything 실험 결과
Fig. 11. Results of Segment-Anything

512x512 또는 1024x1024와 같은 정사각형의 비율을 가진다. 따라서 우리는 Outpainting을 통해 원하는 비율로 조정하고 새로운 영역을 생성한다. 이 과정에서 Stable Diffusion XL의 Outpainting 기술을 사용하여 입력 프롬프트를 기반으로 기존 이미지에서 자연스럽게 이어지는 새로운 이미지를 생성한다. 반면 MaskGIT[23]는 입력 프롬프트 없이도 원본 이미지에 대한 높은 충실도의 Outpainting을 수행하여, 프롬프트에 따른 편향을 줄이고자 한다.

2.3 Image Segmantation

단일 이미지에서 여러 배경과 전경을 구분하기 위해 우리는 Segment Anything[24] 모델을 사용한다. 이 모델은 다양한 Segmentation 기능을 통합하여 1100만 개의 이미지와 11억 개의 마스크로 학습되었으며, 강력한 Zero-shot 성능을 제공한다. 따라서 복잡한 이미지에서도 객체를 정확하게 분할할 수 있다. 그림 11에서 보듯이, Segment-Anything의 데모 페이지를 이용하면 사용자가 마우스를 클릭하여 원하는 영역을 쉽게 분할할 수 있다.

2.4 Image Inpainting

우리는 각 레이어에 대해 Stable Diffusion XL의 Image Inpainting 기술을 사용하여 가려진 부분을 채운다. 이 기술은 원본 이미지, 수정할 부분을 표시한 마스크 이미지, 입력 프롬프트를 이용해 이미지를 수정할 수 있다. 이를 통해 사용자는 가려진 부분을 복구하고, 원하는 객체나 배경 스타일을 추가하는 Inpainting을 수행할 수 있다. 또한, Power-Paint[25]와 Large Mask Inpainting[26] 연구에서는 입력 프롬프트 없이도 마스크 영역을 자연스럽게 채우는 방법을 제안해 프롬프트로 인한 편향을 방지한다. 이렇게 하면 Image Inpainting 과정에서 원본 이미지의 높은 충실도를 유지할 수 있다.

2.5 Animation Generation

우리는 물과 구름 같은 유체의 움직임을 표현하기 위해 생성된 배경 이미지에 애니메이션을 적용하려고 한다. 이를 위해 Eulerian Flow Field[27]를 사용해 시간에 따라 일정한 속도로 변하는 움직임 정보(flow)를 추정한다. 그림 12에서 볼 수 있듯이, 해당 flow 정보를 바탕으로 3D-Cinemagraphy[28] 기술을 활용해 단일 이미지에서 유체의 움직임을 생성하고, 움직임을 적용할 영역을 표시한 마스크와 움직임의 방향 및 속도를 제어할 수 있는 레이블을 제공한다. 이는 사용자가 원하는 애니메이션을 쉽게 적용할 수 있도록 한다. 또한, 후속 연구인 Make-it-4D[29]는 Diffusion 모델을 사용해 Outpainting과 Inpainting을 수행하며, 3D Cinemagraphy보다 큰 카메라 움직임과 긴 비디오를 생성할 수 있는 기능을 제공한다.

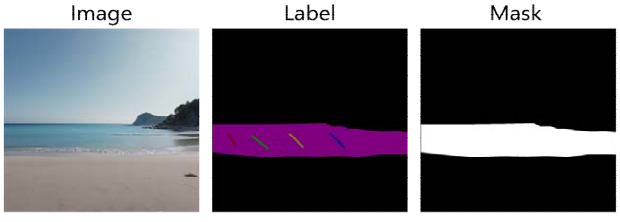

그림 12. 입력 이미지에 대해 Label과 Mask로 적용할 애니메이션에 대한 사용자의 제어 가능성을 제공하고 있는 3D-Cinemagraphy
Fig. 12. 3D-Cinemagraphy providing user control over the animation using labels and masks applied to the input image

우리는 카메라 움직임을 고정시키고 입력 이미지에 유체

의 움직임만 적용하여 결과를 확인한다. 최근 Text2-Cinemagraph[30]는 입력 프롬프트를 통해 생성된 단일 이미지에 애니메이션을 적용하여 예술적인 스타일에서도 인상적인 유체의 움직임을 보여준다. 또한 SLR-SFS[31]는 비디오에서 유체층을 분리해 자연스러운 움직임을 생성한다. 이러한 연구들은 이미지에서 유체의 애니메이션을 적용하는 다양한 방법을 탐구하고 있다.

Ⅳ. 실험 결과

1. 제안하는 알고리즘의 단계별 성능 비교

우리는 제안한 알고리즘 프레임워크에 따라 바다를 배경으로 한 테라스 배경 에셋 생성 실험을 진행하였다.

1.1 Image Generation

Stable Diffusion XL은 사용자가 제공하는 입력 프롬프트를 통해 복잡하고 정교한 이미지를 생성할 수 있으며, 예술적인 스타일의 이미지 등 다양한 형태를 지원한다. 그러나 동일한 입력 프롬프트를 사용하더라도 모델의 자유도에 따라 결과가 달라질 수 있어, 원하는 이미지를 얻기 위해 여러 번의 실험이 필요할 수 있다.

따라서 우리는 ControlNet을 활용해 입력 프롬프트 외에도 추가 정보를 고려하여 더 세밀한 이미지 생성을 지원한다. 그림 13에서 볼 수 있듯이, Stable Diffusion XL로 생성된 이미지는 입력 프롬프트에 충실하지만 테라스의 위치나 크기와 같은 세부적인 요소를 정확히 표현하기는 어려울 수 있다. 이에 따라 사용자가 원하는 배경 에셋의 위치와 크기를 명확히 표시할 수 있도록 스케치를 입력 프롬프트와 결합한다. 스케치는 검은 배경에 흰색 선으로 사용자가 원하는 배경의 크기와 위치를 나타낸다. ControlNet은 스케치를 기반으로 테라스의 위치와 크기, 바다의 위치 등을 일관되게 유지하면서도 입력 프롬프트에 따른 다양한 이미지

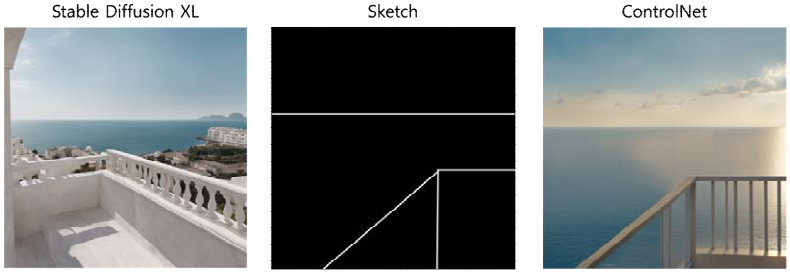

그림 13. 동일한 입력 프롬프트에 대한 Stable Diffusion XL과 ControlNet의 이미지 생성 결과
Fig. 13. Image generation results for the same input prompt using Stable Diffusion XL and ControlNet

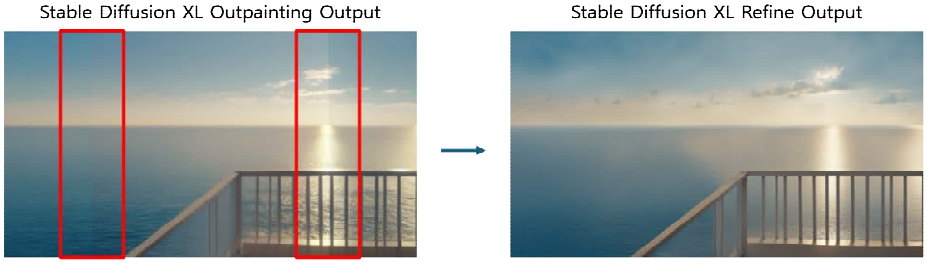

그림 14. Stable Diffusion XL을 통해 Outpainting과 Refine을 수행한 결과
Fig. 14. Results of outpainting and refinement performed using Stable Diffusion XL

결과를 생성할 수 있다.

1.2 Image Outpainting

ControlNet을 통해 생성된 이미지는 512x512 크기로 만들어진 후, 배경 에셋으로 사용하기 위해 이미지의 양 옆을 Outpainting하여 확장한다. 이 과정에서 Stable Diffusion XL의 Outpainting 기술을 활용하며, 입력 프롬프트를 통해 확장할 영역을 정밀하게 조정한다.

그림 14에서 볼 수 있듯이, Outpainting 후 원본 이미지의 양 옆이 자연스럽게 생성되었지만, 확장한 부분의 경계선이 뚜렷하게 남아 있는 문제가 발생한다. 이를 해결하기 위해 우리는 Stable Diffusion XL의 Refine 기능을 도입해 경계선을 자연스럽게 흐리게 한다. 이 과정을 통해 이미지 전반의 노이즈가 감소하고, 구름의 형태나 바다의 물결과 같은 디테일이 개선되었음을 확인할 수 있다.

Stable Diffusion XL로 생성된 이미지에는 무단 사용을 방지하고 저작권을 보호하기 위해 붉은 점 형태의 워터마크가 삽입된다. 이미지를 크게 확대하면 전반적으로 보이는 붉은 색 점이 이에 해당되며 이미지 생성 과정에서 필수적으로 포함되기에 사용자가 임의로 삭제할 수 없다. 따라서 우리는 이미지 Deblur 모델인 NAFnet을 사용하여 워터마크를 제거한다. 그림 15에서 확인할 수 있듯이 Deblur 적용 후 이미지를 확대하면 붉은 점이 사라지고, 동시에 테라스의 경계와 바다의 물결 등이 더욱 선명해져 이미지의 퀄리티가 개선된 것을 확인할 수 있다.

1.3 Image Segmantation

생성된 이미지에 대해 Segmentation을 적용하기 위해 Segment-Anything의 데모 페이지를 활용한다. 이 페이지에서는 사용자가 마우스 클릭으로 분할할 영역을 지정하거나 모든 영역을 자동으로 분할할 수 있다. 우리는 이 기능을 통해 얇은 테라스 기둥의 디테일을 보존하면서 바다와 하늘 배경 레이어, 테라스 레이어를 명확히 구분한다. 그림 16에서 확인할 수 있듯이, 두 개의 분리된 레이어를 통해 카메라 시점에 따라 바다와 하늘, 테라스의 움직이는 속도를 달리하여 배경 에셋에 입체감을 제공한다.

1.4 Image Inpainting

분리된 바다와 하늘 배경 레이어에서는 테라스가 있던 영역에 대한 Inpainting이 필요하다. 이는 카메라 시점에 따라 테라스에 가려졌던 부분이 드러나 새로운 장면이 나타나기 때문이다. 우리는 Stable Diffusion XL의 Inpainting

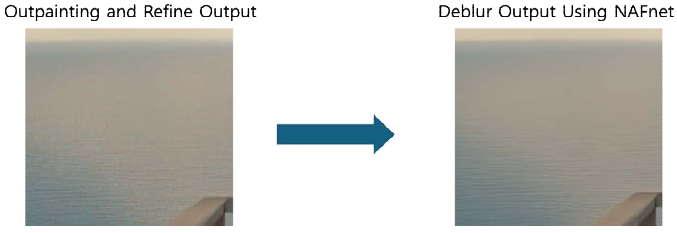

그림 15. NAFnet을 통한 이미지 Deblur 적용 결과
Fig. 15. Results of image deblurring applied using NAFnet

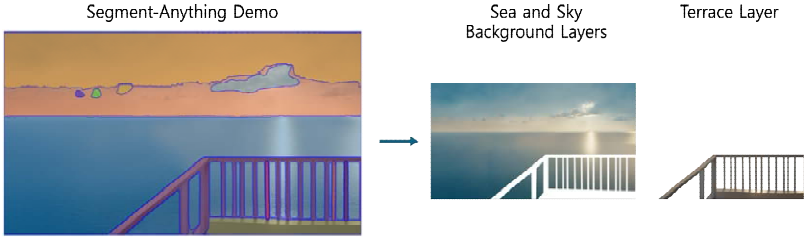

그림 16. Segment-Anything demo 페이지를 통해 segmentation을 수행한 결과
Fig. 16. Results of segmentation performed using the Segment-Anything demo page

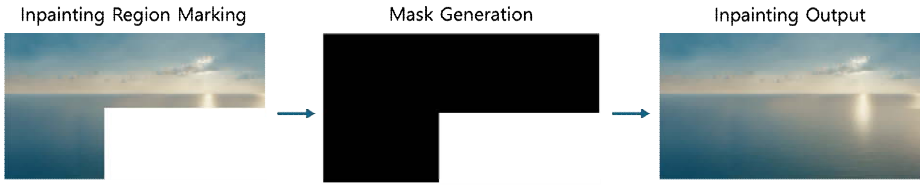

그림 17. Inpainting 영역에 대한 Mask를 생성 후 Stable Diffusion XL 모델을 통해 Inpainting을 수행한 결과
Fig. 17. Results of inpainting performed using the Stable Diffusion XL model after generating a mask for the inpainting area

기능을 사용하여 이 작업을 수행하며, 입력 프롬프트를 통해 생성되는 영역을 세밀하게 조정할 수 있다. Inpainting을 위해서는 입력 프롬프트와 함께 Mask를 필요로 하며, 사용자가 이 영역을 명확히 지정해야 한다. 특히 테라스의 미세한 픽셀이 분리된 바다와 배경 레이어에 영향을 줄 수 있으므로, 우리는 실제보다 더 넓은 영역을 Inpainting하도록 표시한다. 그림 17에서 확인할 수 있듯이, Inpainting되지 않은 영역은 원본 이미지와 동일하며, Inpainting된 영역은 새롭게 생성된 결과를 보여준다. 사용자는 입력 프롬프트를 조정하여 원하는 영역을 세밀하게 제어할 수 있으며, 같은 프롬프트라도 실험마다 다른 결과를 얻을 수 있다.

1.5 Animation Generation

우리는 물과 같은 유체의 움직임을 통해 몰입감 있고 현실적인 배경 에셋 데이터를 생성하고자 한다. 이를 위해 바다와 하늘 배경 레이어에 대해 3D-Cinemagraphy와 Make-it-4D, Text2Cinemagraph, SLR-SFS 모델의 애니메이션 적용 결과를 비교하였다. 애니메이션 비디오 생성에는 움직임이 잘 적용되는 것뿐만 아니라, 움직이지 않는 영역에 원본 이미지의 품질 유지도 중요하다. 그림 18에서 볼 수 있듯이, 애니메이션이 적용되지 않는 부분에서 Text2Cinemagraph의 이미지 품질이 떨어지는 것을 확인할 수 있다. 또한 SLR-SFS와 Make-it-4D는 해상도와 디테일이 부족하다. 반면 3D-Cinemagraphy는 바다의 애니메이션과 움직이지 않는 영역 모두에서 가장 우수한 이미지 품질을 보여준다.

3D-Cinemagraphy는 사용자가 애니메이션이 적용된 영역과 움직임의 방향을 정의하여 더욱 사실적인 움직임을 생성할 수 있다. 또한 원하는 비디오 길이를 조정할 수 있다. 우리는 바다의 잔잔한 물결을 만들기 위해 짧은 물결 방향을 설정하고, 그림 19에서 볼 수 있듯이 60프레임의 비디오를 생성했다. 이 실험에서는 카메라 각도를 고정하여 바다의 애니메이션에 집중했다.

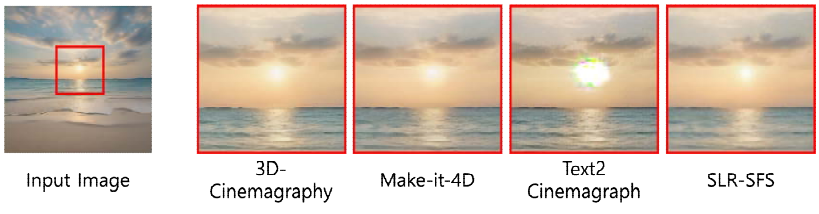

그림 18. 동일한 입력 이미지에 대한 애니메이션 적용 시 이미지 퀄리티 결과
Fig. 18. Image quality results when applying animation to the same input image

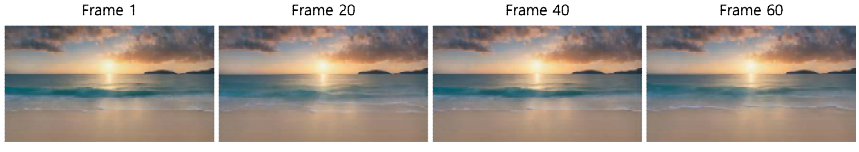

그림 19. 3D-Cinemagraphy를 통한 바다 애니메이션 효과 적용 결과
Fig. 19. Results of applying sea animation effects using 3D-Cinemagraphy

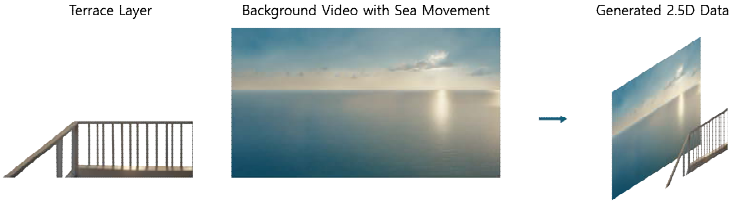

그림 20. 레이어와 비디오를 바탕으로 최종 2.5D 데이터 생성 결과 이미지
Fig. 20. Final 2.5D data generated based on layers and video

1.6 최종 2.5D 데이터 생성

이 과정을 통해 우리는 그림 20에서 볼 수 있는 테라스와 바다의 움직임이 적용된 배경 비디오를 생성하였다. 이를 원본 이미지의 깊이 정보 또는 사용자가 정의하는 깊이를 기반으로 2.5D 데이터로 최종 생성한다. 이를 통해 카메라 각도에 따라 테라스에 가려진 부분이 드러나고, 가까운 테라스와 먼 바다 및 하늘 간의 시차 효과가 적용되어 사용자는 더욱 입체감을 느낄 수 있다. 제공되는 바다는 움직임이 있어 현실적이고 몰입감 있는 경험을 제공한다.

2. 제안하는 알고리즘과 Cuebric의 단계별 비교

우리는 Stable Diffusion XL의 Outpainting과 Inpainting 기술을 사용하여 동일한 입력 이미지에 대해 결과를 생성한다. Cuebric도 입력 이미지와 입력 프롬프트를 이용해 이 과정을 수행한다. 본 논문에서 제안하는 알고리즘과 Cuebric은 공통적으로 Image Segmentation과 Image Inpainting 및 Outpainting을 통해 배경 에셋을 생성하지만, 각 단계에서 결과 이미지에 유의미한 차이가 있어 이를 단계별로 비교하고자 한다.

먼저, Image Segmentation 단계에서는 Cuebric의 'Segment All Object' 기능과 Segment-Anything의 클릭 기능

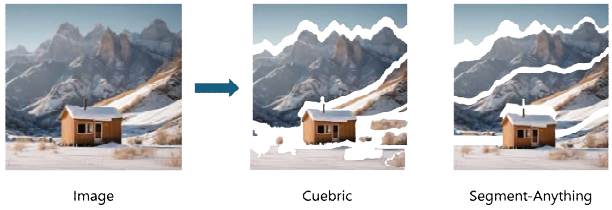

그림 21. 제안하는 모델과 Cuebric과의 Segmentation 결과 비교
Fig. 21. Comparison of segmentation results between the proposed model and Cuebric

을 비교한다. 그림 21에서 볼 수 있듯이, Cuebric의 결과는 깊이 기준 없이 모든 객체를 분리하므로 이후 레이어 생성이 어려운 반면, Segment-Anything은 사용자가 깊이를 기준으로 이미지를 분리했기에 Inpainting 및 Outpainting 작업이 상대적으로 용이하다.

다음 단계에서는 분리된 이미지에서 Inpainting 및 Outpainting을 수행한다. 그림 22의 좌측에는 Inpainting의 결과가 나와 있다. 이 과정에서 하늘과 산을 분할한 영역을 동일한 입력으로 사용했다. Stable Diffusion XL은 구름이 가득한 하늘을 자연스럽게 확장하여 원본의 스타일과 톤을 유지하는 반면 Cuebric의 Inpainting은 거의 구름이 없는 빈 하늘을 생성하여 입력 이미지와 일관성이 떨어지며 경계가 어색하다. 또한, Stable Diffusion XL은 자연스러운 풍경을

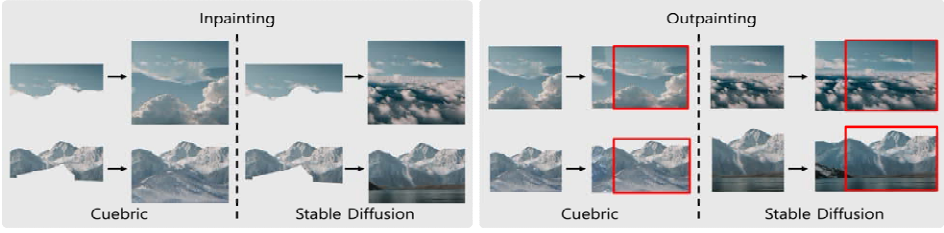

그림 22. 제안하는 모델과 Cuebric과의 Outpainting과 Inpainting 결과 비교
Fig. 22. Comparison of outpainting and inpainting results between the proposed model and Cuebric

생성한 것과는 달리 Cuebric은 산을 늘려서 채워 어색한 시점 이동을 초래할 수 있다.

그림 22의 우측은 Stable Diffusion XL로 생성한 Outpainting 결과이다. 이 결과는 확장된 영역의 경계가 자연스럽고, 생성된 이미지가 원본과 유사한 스타일을 유지한다. 반면, Cuebric의 Outpainting 결과는 경계가 뚜렷하고, 스타일과 톤이 원본과 달라 어색함을 느낄 수 있다.

3. 제안하는 알고리즘의 추가 결과

우리는 제안하는 알고리즘의 우수성을 검증하기 위해 추가 실험을 진행하였으며, 그림 23과 그림 24에서 그 결과를 확인할 수 있다. 그림 23에서는 프롬프트를 통해 생성된 이미지를 총 3개의 레이어로 분리하여 비현실적인 장소에도 현실적인 입체감을 제공하고자 하였다.

그림 23. 제안하는 모델의 추가 결과 1
Fig. 23. Additional results of the proposed model 1

또한 그림 24에서는 프롬프트를 통해 생성된 이미지를 4개의 레이어로 분리하였고, Inpainting 및 Outpainting이 효과적으로 수행되어 사실적인 몰입감을 제공하고 있다. 이처럼 제안하는 알고리즘은 비현실적인 이미지와 현실적인 이미지 모두에서 입체감을 제공할 수 있어, 다양한 배경 에셋을 생성할 가능성을 보여준다.

V. 결 론

본 논문에서는 3D 공간을 효율적으로 표현할 수 있는 배경 에셋 데이터를 생성하는 것을 목표로, 기존 연구인 Cuebric을 검토하고 이를 기반으로 개선된 2.5D 데이터 생성 알고리즘을 제안한다. 2.5D 데이터는 단일 이미지에서 카메라 각도가 변함에 따라 보이지 않던 장면이 나타나고, 사물의 위치에 따라 시차 효과가 적용되어 현실적인 입체감을 제공한다. 또한, 물과 같은 유체의 움직임을 포함하는 풍경을 고려하여, 단일 이미지에서 유체의 움직임을 추정해 애니메이션 비디오를 생성하는 모델을 도입했다. 제안된 알고리즘은 총 5단계로 구성되었으며, 각 단계에서 기존 모델들을 활용하여 개선된 결과를 도출할 수 있었다. 이를 통해 AI를 활용하여 낮은 비용과 시간으로 높은 퀄리티의 배경 에셋 데이터를 생성할 수 있었다. 이러한 접근은 영화 촬영 산업에서도 주목받을 기술로, 에셋 데이터셋의 공급 부족 문제를 효과적으로 해결할 수 있을 것이다.

참 고 문 헌 (References)

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. Generative adversarial nets. Advances in neural information processing systems, 27. 2014. doi: https://doi.org/10.1145/3422622

[2] Sun, C., Hsiao, C. W., Sun, M., & Chen, H. T. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

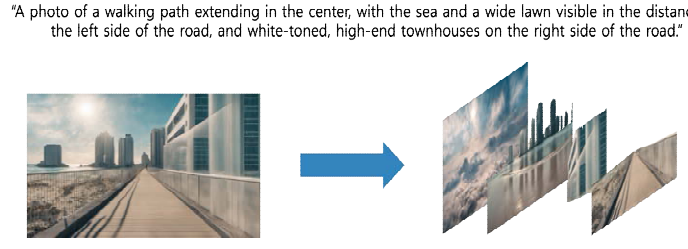

그림 24. 제안하는 모델의 추가 결과 2
Fig. 24. Additional results of the proposed model 2

doi: https://doi.org/10.1109/cvpr.2019.00114

[3] Dong, Y., Fang, C., Bo, L., Dong, Z., & Tan, P. PanoContext-Former: Panoramic total scene understanding with a transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
doi: https://doi.org/10.1109/cvpr.2019.00114

[4] Jia, H., Yi, H., Fujiki, H., Zhang, H., Wang, W., & Odamaki, M. 3d room layout recovery generalizing across manhattan and non-manhattan worlds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
doi: https://doi.org/10.1109/cvprw56347.2022.00567

[5] Höllein, L., Cao, A., Owens, A., Johnson, J., & Nießner, M. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
doi: https://doi.org/10.1109/iccv51070.2023.00727

[6] Deng, Z., He, X., Peng, Y., Zhu, X., & Cheng, L.. MV-Diffusion: Motion-aware video diffusion model. In Proceedings of the 31st ACM International Conference on Multimedia pp. 7255-7263. October, 2023.
doi: https://doi.org/10.1145/3581783.3612405

[7] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM. 2021.
doi: https://doi.org/10.1007/978-3-030-58452-8_24

[8] Cao, A. Q., & de Charette, R. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
doi: https://doi.org/10.1109/iccv51070.2023.00861

[9] Nguyen-Phuoc, T., Liu, F., & Xiao, L. Snerf: stylized neural implicit representations for 3d scenes. ACM Transactions on Graphics (TOG). 2022.
doi: https://doi.org/10.1145/3528223.3530107

[10] Wang, P., Liu, Y., Chen, Z., Liu, L., Liu, Z., Komura, T., ... & Wang, W. F2-nerf: Fast neural radiance field training with free camera trajectories. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.00404

[11] Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. 2023.
doi: https://doi.org/10.1145/3592433

[12] Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF international conference on computer vision. 2021.
doi: https://doi.org/10.1109/iccv48922.2021.00580

[13] Müller, T., Evans, A., Schied, C., & Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG). 2022.
doi: https://doi.org/10.1145/3528223.3530127

[14] Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., ... & Wang, X. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
doi: https://doi.org/10.1109/cvpr52733.2024.00649

[15] Chung, J., Lee, S., Nam, H., Lee, J., & Lee, K. M. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384. 2023.
doi: https://doi.org/10.48550/arXiv.2311.13384

[16] Ouyang, H., Heal, K., Lombardi, S., & Sun, T. Text2immersion: Generative immersive scene with 3d gaussians. arXiv preprint arXiv:2312.09242. 2023.
doi: https://doi.org/10.48550/arXiv.2312.09242

[17] https://cuebric.com/

[18] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 10684-10695. 2022.
doi: https://doi.org/10.1101/2022.11.18.517004

[19] Zhang, L., Rao, A., & Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 3836-3847. 2023.
doi: https://doi.org/10.1109/iccv51070.2023.00355

[20] Zhao, S., Chen, D., Chen, Y. C., Bao, J., Hao, S., Yuan, L., & Wong, K. Y. K. Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems, 36. 2024.
doi: https://doi.org/10.48550/arXiv.2305.16322

[21] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. The Eleventh International Conference on Learning Representations. 2022.
doi: https://doi.org/10.48550/arXiv.2208.01618

[22] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., & Zhu, J. Y. Multi-concept customization of text-to-image diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.00192

[23] Chang, H., Zhang, H., Jiang, L., Liu, C., & Freeman, W. T. Maskgit: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.01103

[24] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
doi: https://doi.org/10.48550/arXiv.2304.02643

[25] Zhuang, J., Zeng, Y., Liu, W., Yuan, C., & Chen, K. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. The 18th European Conference on Computer Vision. 2024.
doi: https://doi.org/10.48550/arXiv.2312.03594

[26] Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., ... & Lempitsky, V. Resolution-robust large mask

inpainting with fourier convolutions. In Proceedings of the IEEE/CVF winter conference on applications of computer vision pp. 2149-2159. 2022.
doi: https://doi.org/10.1109/wacv51458.2022.00323

[27] Holynski, A., Curless, B. L., Seitz, S. M., & Szeliski, R. Animating pictures with eulerian motion fields. In Proceedings of the IEEE/CVF Conference on Computer Vis∟ion and Pattern Recognition. 2021.
doi: https://doi.org/10.1109/cvpr46437.2021.00575

[28] Li, X., Cao, Z., Sun, H., Zhang, J., Xian, K., & Lin, G. 3d cinemagraphy from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.00446

[29] Shen, L., Li, X., Sun, H., Peng, J., Xian, K., Cao, Z., & Lin, G. Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image. In Proceedings of the 31st ACM International Conference on Multimedia. 2023.
doi: https://doi.org/10.1145/3581783.3612033

[30] Mahapatra, A., Siarohin, A., Lee, H. Y., Tulyakov, S., & Zhu, J. Y. Text-guided synthesis of eulerian cinemagraphs. ACM Transactions on Graphics (TOG). 2023.
doi: https://doi.org/10.1145/3618326

[31] Fan, S., Piao, J., Qian, C., Li, H., & Lin, K. Y. Simulating fluids in real-world still images. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
doi: https://doi.org/10.1109/iccv51070.2023.01459

저 자 소 개

추 해 수

- 현재 : 부경대학교 미디어커뮤니케이션학부 휴먼ICT융합전공 학사과정
- ORCID : https://orcid.org/0009-0008-8462-262X
- 주관심분야 : 생성적 적대 신경망, 이미지 처리, 다중 모달성

진 인 환

- 현재 : 부경대학교 미디어커뮤니케이션학부 휴먼ICT융합전공 학사과정
- ORCID : https://orcid.org/0009-0008-9202-6510
- 주관심분야 : 생성적 적대 신경망, 이미지 처리, 다중 모달성

정 성 훈

- 부경대학교 미디어커뮤니케이션학부 언론정보전공 학사
- 부경대학교 미디어커뮤니케이션학과 공학석사
- 현재 : 부산대학교 전기전자공학과 박사과정
- ORCID : https://orcid.org/0000-0002-5505-6189
- 주관심분야 : 생성형 인공지능, 워터마킹, 다중 모달성

저 자 소 개

김 정 환

- 2009년 : 고려대학교 언론학부 학사
- 2011년 : 고려대학교 언론학과 석사
- 2014년 : 고려대학교 언론학과 박사
- 2020년 : 네이버 정책연구실 연구위원
- 현재 : 국립부경대학교 미디어커뮤니케이션학부 휴먼ICT융합전공 부교수
- ORCID : https://orcid.org/0000-0001-5360-0059
- 주관심분야 : 미디어경영, 미디어산업, 엔터테인먼트 테크놀로지

공 경 보

- 2015년 : 서강대학교 전자공학과 학사
- 2017년 : 포항공과대학교 전자전기공학과 석사
- 2020년 : 포항공과대학교 전자전기공학과 공학박사
- 2021년 : 포항공과대학교 전자전기공학과 박사후연구원
- 2023년 : 부경대학교 미디어커뮤니케이션학부 휴먼ICT융합전공 조교수
- 현재 : 부산대학교 전기전자공학부 전자공학전공 조교수
- ORCID : https://orcid.org/0000-0002-1135-7502
- 주관심분야 : 멀티미디어 영상신호처리, 컴퓨터 비전, 딥러닝 시스템 설계