일반논문 (Regular Paper)
방송공학회논문지 제29권 제4호, 2024년 7월 (JBE Vol.29, No.4, July 2024)
https://doi.org/10.5909/JBE.2024.29.4.452
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

NeRF와 이미지 인페인팅을 이용한 가려진 객체의 표면 재구성

강 현 석[a], 박 구 만[a]‡

A Study on Surface Reconstruction of occluded Objects using NeRF and Image Inpainting

HyunSeok Kang[a] and Gooman Park[a]‡

요 약

전통적인 3D 비전 연구에서 가장 대표적인 3D 재구성 방법은 다중 시점 기하학을 이용하여 복원하는 것이었다. 다중 시점 기하학 방법은 대상의 특징점을 검출해야 하므로 특징이 명확한 물체가 있어야 복원이 원활하며, 많은 수의 이미지가 필요하다. 최근 딥러닝의 발전은 이러한 제약조건들을 해결하는 방법들을 제시하고 있다. 특히 3D 재구성에서는 NeRF가 발표된 이후 빠르게 발전하고 있다. 하지만 NeRF 방식은 주어진 카메라 시점을 이용하여 3D를 학습하는 방식이기 때문에, 가려진 부분에 대한 복원은 깨끗하게 이뤄지지 않는다. 본 논문은 이미지 인페인팅 기법을 이용하여 대상의 가려진 부분을 2D 이미지상에서 복원한 뒤, NeRF 학습 과정에서 가려진 부분을 3D로 복원하는 방법을 제안한다. 이를 통해 기존 가려진 이미지를 사용하여 네트워크를 학습했을 때보다 약 46%의 PSNR 성능 향상을 이루었다.

Abstract

In traditional 3D vision research, the most representative 3D reconstruction method is to use Multiple View Geometry. Since the Multiple View Geometry method requires the extract of the target's feature points, it requires clear feature objects and a large number of images for a smooth reconstruction. Recent advances in deep learning have provided ways to address these constraints. In particular, 3D reconstruction has been advancing rapidly since the publication of NeRF. However, since the NeRF method uses a given camera view to learn 3D, it does not restore occluded areas cleanly. In this study, we use the Image Inpainting technique to restore the occluded part of the target on the 2D image, and then restore the occluded part to 3D during the NeRF learning process. This resulted in a PSNR performance improvement of about 46% over training the network using the occluded image.

Keyword : 3D Reconstruction, Neural Radiance Fields, Image Inpainting

a) 서울과학기술대학교 일반대학원 스마트ICT융합공학과(Dept. of Smart ICT Convergence Engineering, Graduate School, Seoul National University of Science&Technology)
‡ Corresponding Author : 박구만(Gooman Park)
E-mail: gmpark@seoultech.ac.kr
Tel: +82-2-970-6430
ORCID: https://orcid.org/0000-0002-7055-5568

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00751, Development of multi-dimensional visualization digital twin framework technology for displaying visible and invisible information with lower than 0.5mm precision).
· Manuscript May 31, 2024; Revised June 5, 2024; Accepted June 5, 2024.

Ⅰ. 서 론

3D 재구성은 2D 이미지로부터 3D를 복원하는 방법이다. 카메라는 3D인 실제 세상을 2D로 투영하여 이미지로 만든다. 이 과정에서 많은 정보가 압축되어 손실되기 때문에, 2D에서 3D를 만드는 것은 쉽지 않다. 따라서 부족한 정보를 보충하기 위해 여러 각도에서 찍은 다양한 이미지를 이용하여 3D를 복원하는 방법인 스테레오 비전이 나오게 되었다. 스테레오 비전을 통해 이미지 사이의 3차원 관계를 설명하는 다중 시점 기하학을 알게 되면, 2D 이미지를 통해 3D를 만들어 낼 수 있다.

최근 딥러닝의 발전으로 2D 분야에서는 매우 큰 변화가 있었다. 이미지 분류, 객체 지역화, 객체 분할 등 다양한 분야에서 기존의 전통적인 컴퓨터 비전 방법보다 딥러닝 방식이 더 성능이 좋다. 하지만 3D 분야에서는 아직 딥러닝 방식들이 기존 연구의 성능을 뛰어넘지 못하고 있다. 3D 재구성 분야에서 데이터는 기본적으로 매우 크기 때문에, 학습에 어려움이 존재한다. 3D 데이터의 예인 복셀, 포인트 클라우드, 메시는 자세한 표현을 할수록 데이터의 크기가 기하급수적으로 커진다.

기존 표현 방법으로 3D를 학습하기 어려워 2D 딥러닝에 비해 3D 딥러닝이 발전하지 못했을 때, 새로운 표현 방법인 암시적 표현(Implicit Representation)이 등장했다. 이는 딥러닝 네트워크가 3D를 명시적으로 학습하는 것이 아닌 암시적으로 학습하는 방법이다. 3D를 복셀, 포인트 클라우드, 메시와 같이 명시적으로 표현하면 해상도에 따라 메모리가 비약적으로 증가하게 되는데 이를 해결하기 위해 3D 포인트 좌표를 입력으로 받아 RGB 색상을 출력하는 네트워크를 학습한다. 이 경우, 기존 딥러닝 네트워크의 출력이 특정 목적의 결과물인 것과 다르게 네트워크 자체가 결과물이 된다. 네트워크가 3D를 표현하는 암시적 표현법이 되는 것이다. 이처럼 네트워크가 3D를 표현하는 방법은 NeRF (Neural Radiance Fields)[1] 연구로부터 시작됐다. NeRF 이후 3D 딥러닝은 매우 빠르게 발전하고 있으며, 후속 연구가 활발한 연구 중 하나다.

NeRF의 훌륭한 연구 성과에도 불구하고, 단점 또한 명확하다. NeRF는 기본적으로 입력 이미지에 의존하여 학습하므로, 이미지에 없는 부분이나 다른 사물에 의해 가려진 영역은 학습하지 못한다. 나무의 열매가 나뭇잎에 의해 가려지듯이 실제 사물은 주변 여러 물건에 의해 가려질 때가 많은데, 그러한 경우 대상을 복원하기 위한 연구는 아직 많이 이뤄지고 있지 않다.

3D 공간을 복원하는 연구는 아니지만, 2D 이미지상에서 가려진 영역을 복원하는 방법은 존재한다. 이러한 연구를 이미지 인페인팅이라 부른다. 이미지 인페인팅은 이미지상에서 가려지지 않은 영역의 정보를 기반으로 가려진 부분의 영역을 채우는 연구이다. 기존에는 가려지지 않은 영역을 주변 특징의 경향을 반영하여 채우는 방법이었지만, 딥러닝의 발전으로 이미지 인페인팅 또한 더 좋은 성능을 보이게 되었다. 특히 생성형 모델은 누락 된 부분을 실제와 유사하게 생성하여 이질감이 적다.

본 연구는 가려진 대상을 완전한 모습의 3D로 복원하는 것을 목표로 한다. 기존 NeRF가 복원하지 못하는 가려진 영역을 복원하기 위해 2D 이미지상에서 이미지 인페인팅 기법을 이용해 가려진 영역을 복원한다. 복원된 2D 이미지를 이용하여 NeRF 알고리즘을 통해 대상을 3D로 복원하여 완전한 3D 결과물을 생성한다. 본 연구를 통해 이전에는 가려진 영역을 3D로 복원할 수 없었던 NeRF의 단점을 보완하여 가려짐이 있음에도 3D 생성을 온전히 할 수 있는 연구에 도움이 될 것으로 보인다.

Ⅱ. 관련 연구

1. NeRF 기반 3D 재구성 알고리즘

NeRF는 입력된 이미지를 학습하여 새로운 시점에서의 이미지를 구하는 시점 합성(View Synthesis) 알고리즘이다. 3D 좌표와 보는 방향을 입력으로 받아 대상의 색상과 밀도를 구하도록 네트워크가 구성되었다. 따라서, NeRF는 대상을 암시적으로 표현한다. NeRF의 학습 방법은 이미지 픽셀에서 Ray를 쏠 때, Ray 위의 3D 포인트들의 색상 합이 이미지 픽셀의 색상과 같다는 아이디어에서 나오게 되었다. 카메라의 원점으로부터 특정 화소 방향으로 Ray가 뻗어나갈 때, 그 이미지 화소의 색상은 Ray가 물체에 닿는 부분을 나타낸다. 따라서 물체의 존재 여부를 측정하는 밀도정보

와 물체의 표면 여부에 대한 가중치인 투명도 정보를 이용하여 모든 Ray 상에 존재하는 포인트들의 가중치 곱이 이미지 픽셀의 색상이 되도록 네트워크를 학습시킨다. 네트워크는 MLP만을 이용하여 구성한다.

NeRF가 이미지만을 사용하여 3D를 학습할 수 있는 훌륭한 성과를 이뤘지만, 3D 표면 재구성을 하기에는 부족한 점이 존재한다. 첫째로, 단 하나의 장면만을 학습하는데도 불구하고 NeRF는 학습에 시간이 굉장히 오래 걸린다. 이는 NeRF가 3D 공간을 학습하기 때문에, 학습에 사용되는 데이터의 수가 많아야 하며, 이러한 데이터는 모두 이미지의 픽셀에서부터 수많은 3D 포인트를 샘플링하여 구하기 때문에 학습이 오래 걸릴 수밖에 없다. 둘째로, 표면을 매끄럽게 학습하지 못한다. NeRF는 샘플링 된 3D 포인트를 학습하는 볼륨 렌더링을 진행하기 때문에 정확한 물체의 표면을 학습하지 않고 물체의 표면이 존재할만한 공간 주변을 학습하므로 표면을 매끄럽게 학습하기 어렵다. 이러한 단점들을 개선하기 위해 후속 연구들이 나오게 되었다.

Instant-NGP[2]는 NeRF의 가장 큰 단점인 느린 학습을 개선하였다. 논문의 아이디어는 인코딩 방식의 개선이다. 기존 NeRF의 Positional Encoding 방식으로 모든 좌표를 계산하게 되면 연산량이 너무 크고, 학습 속도가 느려진다. Instant-NGP는 연산량을 줄이기 위하여 Multi Resolution Hash Encoding 방식을 도입하였다. Multi Resolution Hash Encoding은 다양한 해상도의 격자 레벨(Grid Level)을 정의하고 각 격자 레벨당 하나의 해시 테이블(Hash Table)을 만든다. 이후, 각 해시 테이블값의 보간으로 특정 포인트의 좌표를 인코딩하는 방법이다. 여러 레벨에서 포인트를 인코딩하고 보는 방향과 같은 보조 정보를 결합하여 하나의 포인트 또는 Ray의 특징 벡터로 만든다. 이 특징 벡터를 MLP 망을 통해 학습시키는 것이 Instant-NGP의 네트워크 구조이다. Multi Resolution Hash Encoding을 사용하여 학습할 파라미터를 대폭 줄였으며, 학습 시간 또한 대폭 줄일 수 있었다. 기존 NeRF가 수 시간 학습했던 것에 비해 Instant-NGP는 수 초 만에 학습할 수 있게 되었다.

NeuS[3]는 NeRF를 3D 재구성을 목적으로 개선한 연구이다. NeRF는 시점 합성을 목적으로 만들어진 연구이기 때문에 대상의 3D 복원에는 정밀하지 못하다. NeRF와 같이 물체가 존재하는 공간의 밀도를 학습하는 방법을 볼륨 렌더링 방식이라 부른다. 볼륨 렌더링의 경우 물체의 깊이가 급격히 변하는 부분의 밀도를 찾아내기 쉽지만, 표면을 매끄럽게 찾지 못한다. NeuS는 추가적인 지도 정보 없이 물체의 정확한 표면을 학습하기 위한 연구이다. 볼륨 렌더링 방식의 NeRF와 표면 표현 방식인 Signed Distance Function (이하 SDF)를 결합하여 물체의 표면을 학습하도록 하였다. NeRF는 픽셀의 실제 색상과 Ray로부터 유추한 색상의 차이로 학습하게 된다. 이때, 3차원 물체의 표면이 이미지에 가장 많은 영향을 끼치게 되므로 물체의 밀도와 밀도를 이용하여 만든 투명도를 가중치로 사용하여 Ray의 색상을 유추하게 된다. 따라서 투명도와 밀도의 곱인 가중치 값이 가장 큰 부분이 물체의 표면이 되도록 학습하는 것이 핵심이다. 카메라 중심에서부터 Ray를 따라가다 물체의 표면에서 값이 가장 크고 나머지 영역의 값은 작은 그래프는 종 모양의 그래프와 같이 표현된다. 따라서 가중치의 값이 로지스틱 밀도 분포(Logistic Density Distribution)의 그래프를 따르도록 만들었다. 이를 표현하기 위하여 NeuS는 SDF를 이용하여 밀도를 sigmoid 함수로 정의했다. 밀도와 투명도를 SDF를 이용하여 구했기 때문에 네트워크가 표면에 대한 학습을 더 잘하게 된다. 특히, NeRF에서 사용하는 Color Loss 외에도 SDF를 학습하기 위해 고안된 SDF Loss를 도입하여 표면 주변 포인트들의 SDF 값을 정규화하여 매끄러운 표면을 학습하도록 했다.

2. 이미지 인페인팅 알고리즘

이미지 인페인팅은 이미지의 손상되거나 일부분이 누락된 곳을 복원하는 기술이다. 딥러닝의 발전으로 딥러닝을 이용하여 이미지를 복원하는 기술이 계속되어 연구되고 있으며, 현재는 스마트폰 애플리케이션이나 편집 프로그램에서도 딥러닝을 이용한 이미지 인페인팅 기술을 적용하고 있다. 딥러닝을 이용한 이미지 인페인팅은 크게 CNN 기반의 방법과 생성형 네트워크 기반의 방법으로 나눌 수 있다. CNN 기반 이미지 인페인팅은 오토 인코더 방식으로 누락된 부분을 채운다. 오토 인코더는 인코더를 통해 이미지를 저차원의 잠재 벡터로 바꾼 후, 디코더를 이용하여 잠재 벡터를 원래의 이미지를 만들도록 네트워크가 학습된다. 이미지 인페인팅에서 오토 인코더는 누락된 부분이 포함된 입력 이미지가 아닌,

누락된 부분을 채우도록 잠재 벡터를 학습해야 한다. 따라서, CNN 구조의 네트워크는 누락된 부분을 채우기 위해 이미지의 맥락 정보를 학습하도록 설계되었다. 생성형 네트워크는 딥러닝에서 가장 인기 있는 분야이며, 많은 발전을 이루었다. 네트워크가 이미지를 생성할 수 있다는 것은 많은 이미지를 이해하고 있는 것을 의미하므로, 이미지 인페인팅에서 효과적으로 누락된 부분을 채울 수 있다.

초기 생성형 네트워크는 GAN(Generative Adversarial Network)[4] 방식을 이용하여 연구되었다. GAN은 Generator와 Discriminator 두 개의 망으로 구성되어 있으며, Generator는 실제와 유사한 이미지를 생성하도록 학습하며, Discriminator는 생성된 이미지가 실제 이미지인지, 가짜 이미지인지 판별한다. Generator와 Discriminator는 서로 경쟁적으로 학습하게 된다. GAN을 이용한 이미지 인페인팅은 기존 방법들에 비해 더 사실적으로 복원되었다.

또 다른 생성형 네트워크인 Diffusion[5] 모델은 노이즈로부터 이미지를 생성하는 모델이다. 이를 위해 네트워크는 실제 이미지에 여러 단계에 거쳐 노이즈를 조금씩 섞는 것으로 forward process를 진행하고, 반대로 노이즈로부터 이전 이미지를 복원하는 reverse process를 통해 이미지의 확률적인 분포를 학습함으로써 이미지를 복원한다. Diffusion 모델이 나오면서 생성 모델은 GAN에서 점점 Diffusion으로 넘어가고 있다. 실제로 Diffusion 기반 방법들이 GAN 기반 방법들보다 이미지를 더 다양하고 사실적으로 복원한다. Diffusion의 성능이 잘 나오면서 추가적인 연구로 텍스트를 입력으로 이미지를 생성하는 연구가 진행되었다. 이러한 연구는 텍스트를 컨디션으로 주어 원하는 이미지를 생성할 수 있다는 장점이 있다. Diffusion을 이용한 Text to Image의 대표적인 연구는 Stable Diffusion[6]이 있다. Stable Diffusion의 네트워크는 크게 3가지로 나눌 수 있다. 텍스트를 컨디션 정보인 토큰으로 바꿔주는 CLIP[7]과 토큰을 받아 Diffusion을 통해 생성된 노이즈를 디노이징 하는 U-Net[8], 잠재 벡터를 이미지로 변환해주는 VAE[9]로 구성된다. Stable Diffusion은 Text to Image 연구 중 코드를 공개한 몇 안 되는 연구이다. 코드를 공개한 덕분에 Stable Diffusion을 이용하여 다양한 연구들이 진행되었고, Text to Image 외에 텍스트 가이드 이미지 변환, 이미지 인페인팅, 이미지 아웃페인팅 등의 연구도 진행되었다.

Ⅲ. 제안하는 시스템

본 논문에서 제안하는 가려진 부분을 복원하는 표면 재구성 네트워크는 크게 두 단계로 나뉜다. 첫 번째 단계는 이미지의 가려진 부분을 이미지 인페인팅을 통해 복원하는 것이고, 두 번째 단계는 복원된 이미지를 이용하여 3D 재구성 알고리즘을 학습하는 것이다. 가려진 부분에 의해 3D 학습이 모호해지는 부분을 해결하기 위해 시스템을 구성하였다.

1. 이미지 인페인팅 파이프라인

첫 번째 단계는 이미지의 가려진 부분을 이미지 인페인팅을 통해 복원하는 과정이다. 그림 1의 Multiple View Images는 여러 각도에서 찍은 이미지 데이터 세트를 구성

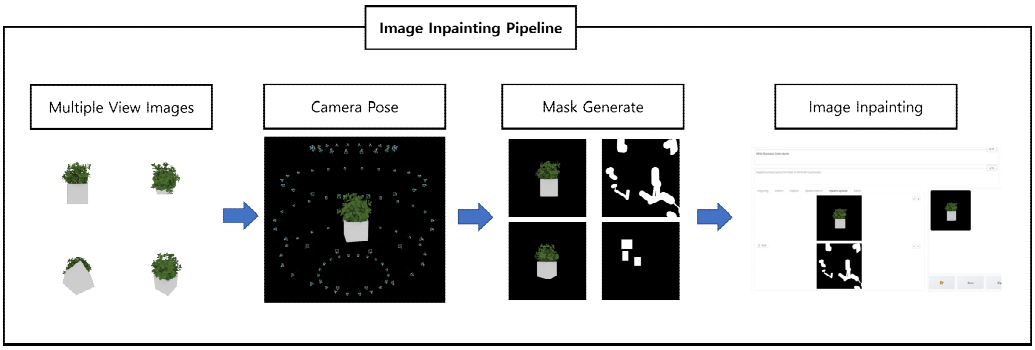

그림 1. 제안하는 시스템의 이미지 인페인팅 파이프라인
Fig. 1. Image inpainting pipeline of the proposed system

하는 것이다. 이후 Camera Pose에서는 여러 각도에서 찍은 이미지에서 카메라 포즈를 구한다. 카메라 포즈를 구하는 방식은 이미지 데이터 세트마다 차이가 있으며, 3D 프로그램으로 만든 데이터 세트의 경우 프로그램에서 제공해주는 포즈 정보를 취득하고, 실제로 촬영한 이미지 데이터 세트의 경우 COLMAP과 같은 Structure from Motion 프로그램을 이용해 카메라 포즈 정보를 취득한다.

Mask Generate는 모든 이미지에서 가려진 영역을 생성하기 위해 마스크 이미지를 생성한다. 마스크 이미지는 그림 2와 같이 랜덤한 위치에 원, 선, 사각형 등의 모양을 랜덤으로 섞어서 생성한다. 마스크 이미지를 이용하여 원본 이미지에서 마스크 이미지를 뺀 가려진 이미지를 생성한다. 마지막으로 Image Inpainting에서는 각 이미지의 마스크로 인해 가려진 영역을 이미지 인페인팅 알고리즘을 이용해 가려진 영역을 복원한다. 이미지 인페인팅 알고리즘은 Stable Diffusion을 사용한다. Stable Diffusion은 이미지 생성에 특화된 생성형 인공지능 모델이며, 텍스트로 이미지

그림 2. 마스크 이미지를 이용하여 가려진 이미지 생성
Fig. 2. Create occlusion images using mask images

생성, 이미지 변환, 이미지 인페인팅, 이미지 아웃페인팅 등의 기능을 사용할 수 있다. Stable Diffusion이 다른 이미지 인페인팅 알고리즘보다 좋은 이유 중 하나는 텍스트를 가이드로 줄 수 있다는 점이다. 텍스트를 가이드로 주어 일반적인 이미지 인페인팅 알고리즘들이 복원하지 못했던 세부 정보들을 복원할 수 있게 한다.

그림 3의 예시에서 Stable Diffusion으로 인페인팅을 진행했을 때, 의자의 모습이 그럴듯하게 복원되었지만, 의자의 팔걸이와 같은 세부적인 부분은 잘 복원하지 못한다. 텍스트 가이드로 'Chair, Beige and Green'과 같은 단어를 주어 인페인팅을 진행하자 원래 복원하지 못했던 팔걸이 부분의 초록색 받침과 의자 시트 부분이 더 잘 복원된 것을 확인할 수 있었다. 따라서 Stable Diffusion을 사용해 텍스트 가이드 된 이미지 인페인팅을 하여 이미지를 복원하며, 텍스트는 각 이미지마다의 특징을 적는다. Stable Diffusion의 파라미터 조절에서 이미지의 리사이즈는 진행하지 않으며 스케일을 1로 둔다. 마스킹 된 부분만 인페인팅을 진행하도록 Mask Mode를 Inpaint masked, Inpainted area를 Only masked로 설정한다. 이러한 설정의 이유는 Stable Diffusion이 전체 이미지를 인페인팅 한다면 누락된 부분을 복원했을 때, 전체 이미지가 어색하지 않게 복원하기 때문에 누락되지 않은 부분도 변형이 일어나기 때문이다. 위와 같은 방법으로 이미지 인페인팅을 진행하여 객체의 일부가 가려진 이미지 데이터 세트를 구성하였다.

2. 3D 재구성 파이프라인

두 번째 단계는 복원된 이미지를 이용하여 NeRF 기반의 네트워크를 학습한다. 그림 4의 Multiple Inpainting Images

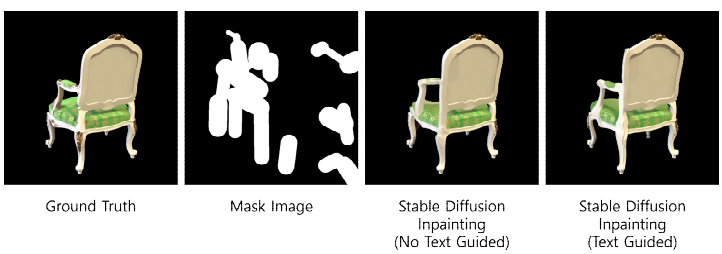

그림 3. Stable Diffusion의 텍스트 가이드 이미지 인페인팅 비교
Fig. 3. Comparison of Stable Diffusion's text guided image inpainting

는 앞서 이미지 인페인팅 파이프라인을 통해 구한 인페인팅 이미지와 카메라 포즈 정보를 데이터 세트로 갖는다. Training Networks는 인페인팅된 이미지를 이용하여 3D 딥러닝 신경망을 학습하는 과정이다. 3D를 학습할 네트워크는 NeuS 기반의 신경망이며, 인코딩 방식은 Instant-NGP의 Multi Resolution Hash Encoding 방법을 사용한다. NeuS는 RGB 색상을 구하기 위한 네트워크와 SDF를 구하기 위한 네트워크가 연결된 형태의 네트워크다. 네트워크는 단순한 MLP의 연속으로 구성되어 있으며, hidden layer의 차원은 64이다. 네트워크의 입력은 Hash Encoding을 통과시킨 3D 포인트(x, y, z)이며, 이를 통해 SDF 값과 특징 벡터를 얻게 되고, RGB 색상은 특징 벡터와 Hash Encoding을 통과시킨 보는 방향 정보, 법선 벡터 값을 입력으로 구한다. MLP는 모두 ReLU를 활성 함수로 가지며 마지막에 RGB를 구할 때는 Sigmoid를 활성 함수로 갖는다.

네트워크의 파라미터는 각 Ray의 샘플 수는 1,024개, 배치 당 훈련 ray 수는 256개, 최대 훈련 ray 수는 8,192개로 정의한다. 네트워크의 훈련은 두 단계에 걸쳐서 진행한다. 두 단계에 걸쳐서 학습을 진행하는 이유는, 인페인팅 이미지만 이용하여 학습하는 것이 Ray의 밀도를 잘 학습하지 못하기 때문이다. 이는 인페인팅 된 포인트가 여러 이미지에서 일관되게 같은 색상을 갖지 못하기 때문에 Ray가 학습을 잘하지 못하는 것으로 보인다. 이를 해결하기 위하여 네트워크의 훈련을 두 단계로 나누어 진행한다. 첫 번째 단계에서는 인페인팅이 되지 않은 화소에서 Ray를 학습한다. 정확한 색상으로부터 네트워크를 학습하기 때문에, 비어있는 영역이 존재하지만, 밀도 학습을 잘하게 된다. 두 번째 단계에서는 인페인팅 된 화소에서 Ray를 학습한다. 첫 번

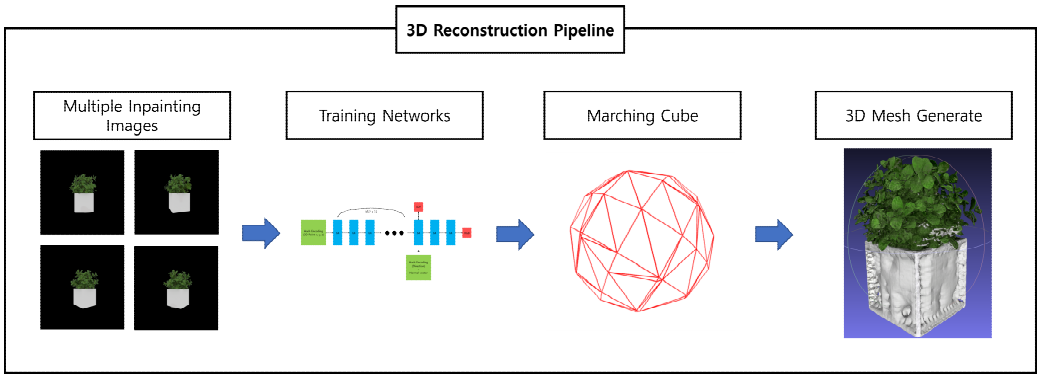

그림 4. 제안하는 시스템의 3D 재구성 파이프라인
Fig. 4. 3D Reconstruction pipeline of the proposed system

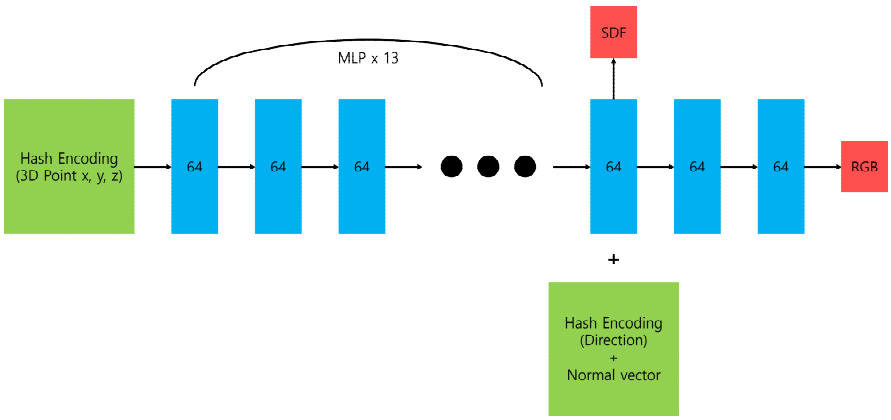

그림 5. 본 연구에서 제안하는 시스템의 표면 재구성 네트워크 구조
Fig. 5. The surface reconstruction network structure of the system proposed in this study

째 단계에서 네트워크가 정확한 밀도를 가지도록 Ray를 학습했기 때문에 인페인팅을 통해 생성된 화소를 이용하여 학습하였을 때, 일정 수준 이상의 학습이 가능하다. 첫 번째 단계에서는 정상적인 Ray 학습을 위해 MSE Loss를 10.0, Eikonal Loss를 0.1의 가중치로 주어 Loss를 구성하고, 학습률(Learning Rate)은 0.01로 설정하여 학습을 진행한다. 충분한 학습을 위하여 20,000 Iteration 학습한다. 두 번째 단계에서는 첫 번째 단계에서 잘 학습된 색상과 밀도를 기반으로 인페인팅된 이미지의 Ray를 학습한다. MSE Loss를 10.0, Eikonal Loss를 10.0, Opaque Loss를 10.0으로 가중치를 준다. 또한, 이미 학습이 되어있는 상태에서 인페인팅 이미지를 학습하는 것이기 때문에, 학습률을 0.0005로 설정하고 1,000 Iteration 학습한다.

네트워크가 3D를 학습하였다고 해서 바로 3D를 생성할 수 있는 것은 아니다. 왜냐하면, NeRF 기반 네트워크들은 암시적 학습법이기 때문에 결과물이 3D가 아니기 때문이다. NeRF 기반 네트워크의 결과물은 색상과 밀도가 존재하는데, 메시를 생성하기 위해서 밀도정보를 이용한다. 그림 4의 Marching Cube 과정과 같이 3D 메시를 만드는 방법은 마칭 큐브(Marching Cube)[10]라 불리는 알고리즘을 사용하여 만든다. 마칭 큐브란 3차원 복셀 격자에 물체가 존재한다고 가정하고, 물체가 점유하는 영역에 맞춰 큐브를 잘라 물체의 표면에 가깝게 메시를 만드는 알고리즘이다. 따라서 NeRF의 결과물을 이용하여 3D 메시를 만드는 과정은 특정 격자의 큐브에서 일정 3D 포인트마다 밀도를 구하여, 임곗값 이상의 포인트를 점유하면 큐브를 깎는다. 마칭 큐브의 해상도를 늘릴수록 3D 메시를 더 정확하게 구할 수 있지만, 해상도가 커질수록 필요한 메모리도 커지기 때문에 모든 데이터의 마칭 큐브 해상도는 512x512x512로 정의하였다. 마지막으로 3D Mesh Generate 과정에서 텍스쳐를 입힌 3D 메시를 만든다.

Ⅳ. 실험 및 분석

1. 실험 환경

실험은 이미지 인페인팅을 통해 누락된 부분을 복원한 이미지들로 NeRF 기반의 알고리즘을 학습시켜 시점 합성의 성능을 비교한다. 이후, 훈련된 네트워크로부터 3D 재구성을 수행하고 생성된 3D 메시가 얼마나 잘 만들어졌는지 비교한다. 딥러닝 프레임워크는 pytorch를 사용하였으며, pytorch lightning을 이용해 코드를 단순화하였다. NeuS 학습망은 nerfacc api를 이용하여 작성되었다. nerfacc는 NeRF 학습을 위한 pytorch 기반의 python api이다. nerfacc를 이용하여 NeRF, NeuS, Instant-NGP 기반의 네트워크를 최적화하며, 효율적으로 학습할 수 있다.

표 1. 실험환경
Table 1. Experimental environments

H/W	CPU	Intel Core i9-7900X
	GPU	NVIDIA RTX A6000 48GB
	RAM	DDR4 64GB
S/W	OS	Ubuntu 18.04
	CUDA	cuda 11.6
	Python	python 3.8.16
	Pytorch	pytorch 1.13.1

2. 평가 지표

본 연구는 NeRF 기반 네트워크를 사용하여 3D 복원을 진행할 때 가려진 영역에 대해 2D 인페인팅 기법을 적용하여 가려진 영역도 잘 복원하는 것이 목적이므로, PSNR 평가 지표를 사용한다. PSNR(Peak Signal-to-Noise Ratio)은 영상의 손실 정보를 평가할 때 사용되는 평가 지표이다. 신호가 가질 수 있는 최대 전력에 대한 잡음의 전력이기 때문에 영상의 손실이 얼마나 되는지 평가할 수 있다. PSNR 식은 (1)과 같다.

$$PSNR = 10 \cdot log_{10}(\frac{MAX_I^2}{MSE}) \quad (1)$$

추가적으로, 본 연구는 딥러닝을 이용하여 영상을 복원하기 때문에 PSNR뿐만 아니라 FID와 LPIPS를 이용하여 생성된 이미지의 정확도를 확인하였다.

3. 데이터 세트

데이터 세트는 NeRF 학습을 위한 여러 각도에서 찍은

이미지와 각 이미지의 카메라 포즈, 그리고 이미지 인페인팅에서 사용할 이미지의 가려진 부분을 표시하는 마스크 이미지를 준비한다. 학습에 사용할 데이터는 NeRF Synthetic 데이터셋과 Custom 데이터셋 두 가지를 사용한다.

NeRF Synthetic 데이터셋은 NeRF 논문에서 제안한 데이터셋으로, Blender를 이용하여 만든 3D 데이터를 사용해 여러 각도에서 이미지와 카메라 포즈를 구한 데이터셋이다. Blender와 같은 3D 프로그램은 3D 객체를 정교하게 만들 수 있을 뿐 아니라 카메라 포즈 또한 편하게 구할 수 있다. 프로그램상에서 정보들을 다 가지고 있기 때문이다. 원하는 카메라 위치에서 렌더링을 진행하여 이미지를 획득한다. 이미지의 해상도는 800x800이며, 총 8개의 장면에서 각각 훈련 이미지 100장, 검증 이미지 100장, 테스트 이미지 200장으로 구성되어 있다. 또한, 테스트 이미지는 RGB 이미지뿐 아니라 깊이 이미지와 법선 이미지까지 존재한다.

커스텀 데이터셋은 NeRF Synthetic 데이터셋과 유사하게 3D 프로그램을 이용하여 구성하였다. 3D 프로그램을 이용하여 데이터 세트를 구성할 때 실제 이미지를 사용하는 것보다 노이즈나 빛에 의한 영향을 적게 받는다는 장점이 있다. 또한, 이미지를 렌더링하였을 때 광원의 위치나 이미지의 해상도, 카메라의 위치들을 자유롭게 정할 수 있어 이미지 생성에 유리하다. 마지막으로 이미지를 구할 때 배경 없이 구할 수 있다는 것도 큰 장점이다. 3D 재구성에서 복원하고자 하는 대상 외의 정보들은 노이즈가 될 수 있기 때문이다. 본 연구에서는 실험을 위해 식물 모형의 3D 객체를 사용하여 데이터를 취득했다. 이미지의 해상도는

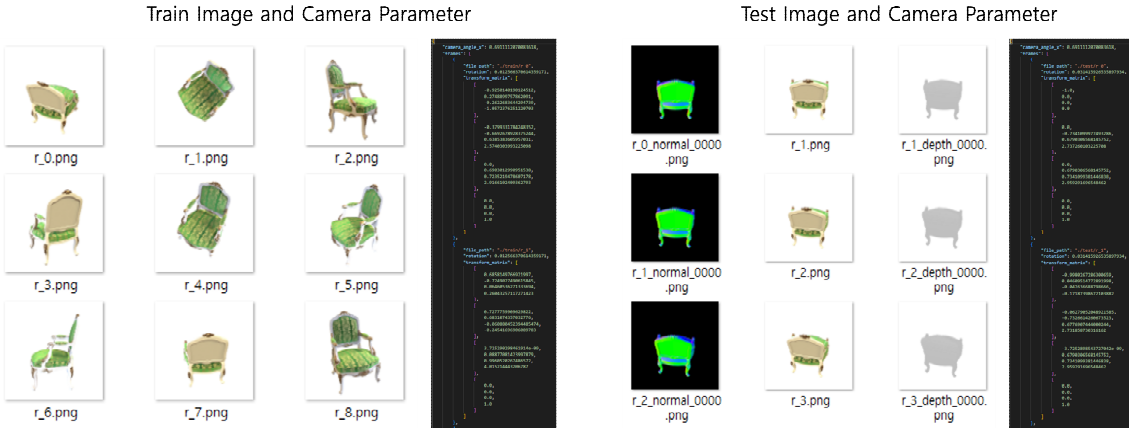

그림 6. 데이터셋의 train, test 이미지와 파라미터
Fig. 6. Dataset of train and test image and parameter

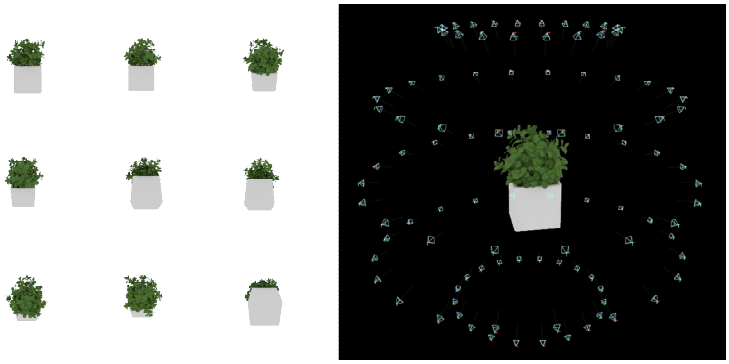

그림 7. 커스텀 데이터셋의 이미지와 카메라 포즈
Fig. 7. Image and camera pose of custom dataset

1024x1024이며, 5개 높이에서 총 100장의 이미지를 취득하였다.

4. 실험 결과 및 분석

표면 재구성 네트워크 학습 결과는 다음과 같다. 그림 8은 원본 이미지, 그림 9는 마스크에 의해 가려진 이미지, 그림 10은 이미지 인페인팅 파이프라인을 통해 복원된 이미지를 이용해 3D 재구성 네트워크를 학습시켜 출력한 결과들이다.

그림 8, 9, 10에서 표면 재구성 네트워크의 결과는 왼쪽부터 원본 이미지, 네트워크가 출력한 RGB 이미지, 깊이 이미지, 법선 벡터 이미지 등이 있다. 실험 결과 마스크에 의해 가려진 이미지를 학습하자 그림 9와 같이 RGB, 깊이, 법선 벡터 이미지 어느 것도 제대로 나오지 않았다. 하지만 이미지 인페인팅을 거쳐서 학습한 경우, 그림 10과 같이 가려진 이미지를 학습했던 것보다 훨씬 깨끗한 RGB, 깊이, 법선 벡터 이미지를 구할 수 있었다. 이미지 인페인팅을 통해 학습한 네트워크의 결과물이 마스크에 의해 가려진 이미지를 학습한 네트워크의 결과물보다 좋지만, 실제 이미지와 비교하였을 때 RGB, 깊이, 법선 벡터 이미지 모두 흐리거나 노이즈가 낀 채로 나오게 되었다. 또한, 데이터셋에 따라 데이터 생성에 차이가 발생하였다. NeRF Synthetic 데이터인 Chair와 Lego의 경우 RGB, 깊이, 법선 벡터 모두 실제 이미지 학습과 인페인팅 이미지 학습 모두 좋은 성능을 보였지만, 커스텀 데이터셋인 Plant 데이터의 경우 성능이 좋지 못하였다. Plant 데이터는 실제 이미지로 학습한 때도 노이즈가 많이 생겼다. 그림 11은 원본 이미지를 학습한

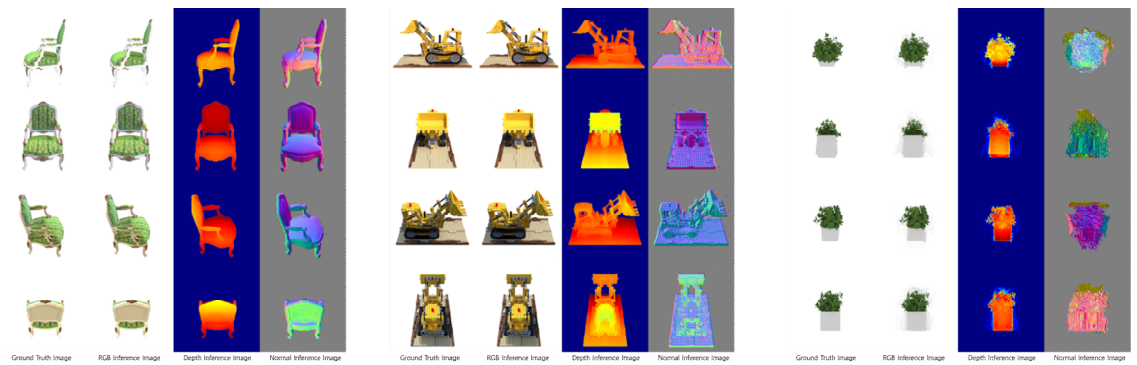

그림 8. 원본 이미지 학습 결과
Fig. 8. Ground Truth image train result

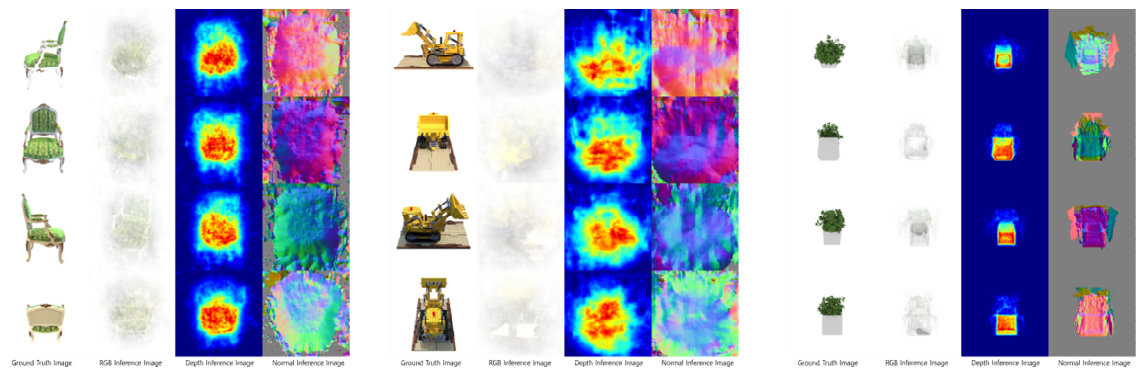

그림 9. 가려진 이미지 학습 결과
Fig. 9. Masked image train result

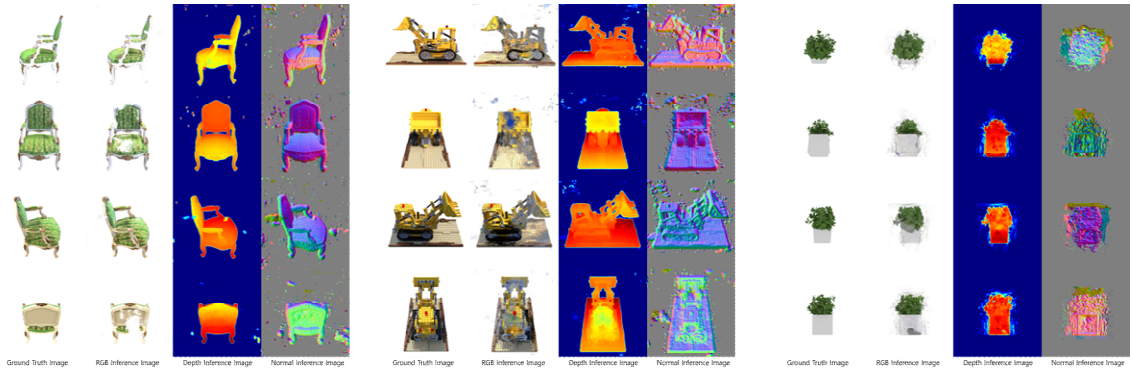

그림 10. 인페인팅 이미지 학습 결과
Fig. 10. Inpainting image train result

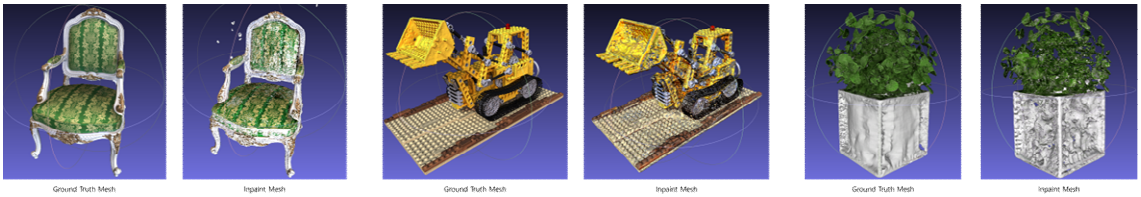

그림 11. 메시 생성 결과
Fig. 11. Mesh generate result

네트워크와 인페인팅 이미지를 학습한 네트워크를 이용해 메시를 생성한 결과이다. 메시 생성은 실제 이미지를 이용하여 학습하였을 때 거의 정확하게 나오는 것을 확인할 수 있었다. 특히 NeRF Synthetic 데이터셋인 Chair와 Lego를 이용하여 만든 메시는 실제 3D와 거의 유사하였으며, 자체 데이터셋인 Plant의 경우 실제 이미지를 사용했다 하더라도 표면이 매끄럽지 않았다. 마스킹된 이미지에서는 메시가 아예 생성되지 않았으며, 인페인팅된 이미지를 학습한 결과 메시가 실제 이미지를 이용하여 학습하였을 때와 유사하게 생성이 되었다. 다만, 실제 이미지를 이용하여 학습했을 때보다 노이즈가 많이 발생하였으며, 표면이 매끄럽지 않았다.

표 2는 Ground Truth 이미지와 마스크 이미지, 인페인팅 이미지를 학습한 네트워크의 PSNR 수치이다. 제안한 방법을 이용해 인페인팅 이미지를 학습한 네트워크의 PSNR 수치는 Ground Truth 이미지를 학습한 네트워크에 비해 수치가 다소 떨어지지만, 마스크 이미지를 학습한 네트워크에 비해 평균적으로 46.22% 높은 수치를 보였다. 표 3은 train 데이터에서 SPIn-NeRF[12]와 FID 및 LPIPS 수치를 비교한 표이다. SPIn-NeRF는 이미지 인페인팅과 NeRF를 결합한 연구 중 하나이다. 이미지에서 마스크로 선택한 부분을 인페인팅한 뒤 NeRF로 학습하는 것은 본 논문에서 제안한 방법과 유사하다. PSNR은 평균적으로 제안한 방법에 비해 67%의 성능을 보여 나쁘지 않은 결과를 보이지만, 실제 생성된 이미지는 그림 12와 같이 완전히 다른 이미지가 나오는 것을 확인할 수 있다. 이는 FID와 LPIPS를 통해서도 확인할 수 있다.

표 2. PSNR 스코어
Table 2. PSNR score

Dataset (test)	Ground Truth Image	Masked Image	Inpainting Image (Ours)
NeRF Synthetic Chair	34.76190	15.41589	20.28827
NeRF Synthetic Lego	33.17179	11.05949	18.84393
Custom Plant	29.41272	17.57226	24.03677

표 3. SPIn-NeRF와 FID 및 LPIPS 스코어 비교
Table 3. FID and LPIPS score Compare with SPIn-NeRF

Dataset (train)	Inpainting Image (Ours)			SPIn-NeRF		
	PSNR↑	FID↓	LPIPS↓	PSNR↑	FID↓	LPIPS↓
NeRF Synthetic Chair	41.55290	87.33173	0.56577	27.98797	521.97825	0.88287
NeRF Synthetic Lego	40.23199	123.23349	0.50261	27.81370	563.96667	0.83537
Custom Plant	43.31239	182.47191	0.61861	27.98919	505.67685	0.94715

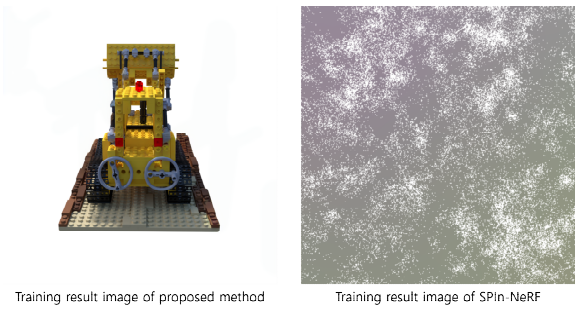

그림 12. 제안한 방법과 SPIn-NeRF의 이미지 비교
Fig. 12. Image comparison of SPIn-NeRF and proposed method

V. 결 론

본 논문에서는 가려진 영역에 대한 표면 재구성을 하는 방법에 대해 실험을 진행하였다. 가려진 영역을 복원하기 위해 이미지 인페인팅 기법을 적용하였으며, 사용한 알고리즘은 생성형 모델인 Stable Diffusion이다. 인페인팅 진행 시 표현이 복잡한 부분은 텍스트를 가이드로 주어 가려진 영역을 생성하였으며 인페인팅을 완료한 데이터를 이용하여 표면 재구성을 진행하였다. 표면 재구성 학습망은 NeuS와 Multi Resolution Hash Encoding 방식을 이용하였다. 실험 결과, 이미지 인페인팅을 통해 복원한 이미지를 이용하여 3D 재구성을 진행하자 가려진 이미지를 이용하여 재구성을 진행할 때보다 평균 약 46%의 PSNR 성능 향상을 보였다. 또한, 기존 인페인팅을 이용한 NeRF 학습 논문인 SPIn-NeRF와 성능 비교를 진행하였다. SPIn-NeRF는 마스크 된 영역을 지운 후 전체 이미지의 맥락을 이용하여 인페인팅하는 방법으로 학습되기 때문에 본 논문에서 사용한 데이터로 학습이 잘 안된 것을 알 수 있다. PSNR은 다소 높게 나왔지만, 실제 생성된 이미지는 원본 이미지와 다르며, 이는 FID와 LPIPS로 확인할 수 있었다. 부정확한 인페인팅으로 인해 초기에는 메시 생성이 불완전했으나, 훈련을 두 단계로 나누어 인페인팅 되지 않은 픽셀의 Ray를 우선 학습하고, 인페인팅 된 픽셀의 Ray를 학습하는 것으로 원본과 유사한 메시 생성을 할 수 있었다. 현재 이미지 인페인팅과 NeRF를 연결한 논문은 대부분 이미지에서 특정 객체를 지우고, 지워진 영역을 어색하지 않게 인페인팅 하여 네트워크를 학습한다. 본 논문은 객체를 지우고 배경을 인페인팅 하는 기존 논문들의 방향과 다르게 가려진 객체를 복원한다는 점에서 차이점이 있다. 향후 위 연구를 발전시킴으로써 객체 복원을 중점으로 한 연구가 더 많아질 것으로 기대한다.

참 고 문 헌 (References)

[1] Ben Mildenhall, Pratul P. Srinivasan, “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”, The European Conference on Computer Vision, Glasgow, UK. 2020.
doi: https://doi.org/10.1007/978-3-030-58452-8_24

[2] Thomas Muller, Alex Evans, Christoph Schied, Alexander Keller, “Instant neural graphics primitives with a multiresolution hash encoding”, ACM Transactions on Graphics, Vol 41, Issue 4, pp.1-15, 2022.
doi: https://doi.org/10.1145/3528223.3530127

[3] Peng Wang, Lingjie Liu, Yuan Liu, “NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction”, Conference on Neural Information Processing Systems, 2021.

doi: https://doi.org/10.48550/arXiv.2106.10689

[4] Ian J. Goodfellow, Jean Pouget-Abadie, “Generative Adversarial Nets”, Conference on Neural Information Processing Systems, 2014.
doi: https://doi.org/10.48550/arXiv.1406.2661

[5] Jonathan Ho, Ajay Jain, Pieter Abbeel, “Denoising Diffusion Probabilistic Models”, Advances in Neural Information Processing Systems, pp.6840-6851, 2020.
doi: https://doi.org/10.48550/arXiv.2006.11239

[6] Robin Rombach, Andreas Blattmann. “High-Resolution Image Synthesis with Latent Diffusion Models”, Conference on Computer Vision and Pattern Recognition, New Orleans, USA, pp.10684-10695, 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.01042

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, “Learning Transferable Visual Models From Natural Language Supervision”, International Conference on Machine Learning, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[8] Olaf Ronneberger, Philipp Fischer, Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, Medical Image Computing and Computer-Assisted Intervention, Vol. 9351, pp. 234-241, 2015.
doi: https://doi.org/10.48550/arXiv.1505.04597

[9] Diederick P Kingma, Max Welling, “Auto-Encoding Variational Bayes”, The International Conference on Learning Representations, AB, Canada, 2014.
doi: https://doi.org/10.48550/arXiv.1312.6114

[10] William E. Lorensen, “Marching cube: A high resolution 3D surface construction algorithm”, Conference on Computer Graphics and Interactive Techniques, California, USA, Vol. 21, Issue 4, pp.163-169, 1987.
doi: https://doi.org/10.1145/37402.37422

[11] Cignoni, Paolo, Callieri, “MeshLab: an Open-Source Mesh Processing Tool”, Eurographics Italian Chapter Conference, Salerno, Italy, pp.129-136, 2008.
doi: https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136

[12] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, “SPIn-NeRF: Multiview Segmentation and Perceptual Inpainting with Neural Radiance Fields”, Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, pp.20669-20679, 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.01980

저 자 소 개

강 현 석

- 2015년 3월 ~ 2021년 8월 : 강원대학교 삼척캠퍼스 전자공학과 학사
- 2021년 9월 ~ 2023년 8월 : 서울과학기술대학교 일반대학원 스마트ICT융합공학과 석사
- 2023년 9월 ~ 현재 : 서울과학기술대학교 일반대학원 스마트ICT융합공학과 박사과정
- ORCID : https://orcid.org/0000-0003-0783-3841
- 주관심분야 : 3D 컴퓨터 비전, 딥러닝

박 구 만

- 1984년 : 한국항공대학교 전자공학과 공학사
- 1986년 : 연세대학교 대학원 전자공학과 석사
- 1991년 : 연세대학교 대학원 전자공학과 박사
- 1991년 ~ 1996년 : 삼성전자 신호처리연구소 선임연구원
- 1999년 ~ 현재 : 서울과학기술대학교 스마트ICT융합공학과 교수
- 2006년 ~ 2007년 : Georgia Institute of Technology, Dept. of ECE. Visiting Scholar
- 2016년 ~ 2017년 : 서울과학기술대학교 나노IT디자인융합대학원 원장
- ORCID : https://orcid.org/0000-0002-7055-5568
- 주관심분야 : 컴퓨터 비전, 실감미디어