일반논문 (Regular Paper)
방송공학회논문지 제28권 제1호, 2023년 1월 (JBE Vol.28, No.1, January 2023)
https://doi.org/10.5909/JBE.2023.28.1.109
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

2D to 3D 창의적 생성을 위한 탐색적 실험 분석

조 형 래[a)], 장 일 식[a)], 강 현 석[a)], 고 영 찬[a)], 박 구 만[a)‡]

Exploratory Experimental Analysis for 2D to 3D Generation

Hyeongrae Cho[a)], Ilsik Chang[a)], Hyunseok Kang[a)], Youngchan Go[a)], and Gooman Park[a)‡]

요 약

딥러닝은 최근 몇 년 동안 비약적인 발전을 하였고 다양한 분야 및 산업에 영향을 주고 있다. 예술영역도 예외일 수는 없는데 본 논문에서는 시각예술·공학적 관점에서 2D 이미지를 3D로 창의적으로 생성하는 방법을 실험하고자 한다. 이를 위해 국내 아티스트 원본 이미지를 GAN 또는 Diffusion Models로 학습시킨 후 3D 변환 소프트웨어와 딥러닝을 활용하여 3D로 변환하고 그 결과를 선행연구 알고리즘과 비교 실험함으로써 2D to 3D 창의적 생성의 문제점과 개선점을 분석하고자 한다.

Abstract

Deep learning has made rapid progress in recent years and is affecting various fields and industries. The art field cannot be an exception, and in this paper, we would like to explore and experiment and analyze research fields that creatively generate 2D images in 3D from a visual arts and engineering perspective. To this end, the original image of the domestic artist is learned through GAN or Diffusion Models, and then converted into 3D using 3D conversion software and deep learning. And we compare the results with prior algorithms. After that, we will analyze the problems and improvements of 2D to 3D creative generation.

Keyword : 2D to 3D , Generation, Point to Mesh, 3D Aware GAN, Creative Deep Learning

a) 서울과학기술대학교 전자IT미디어공학과(Department of Electronic IT Media Engineering, Seoul National University of Science and Technology)
‡ Corresponding Author : 박구만(Gooman Park)
E-mail: gmpark@seoultech.ac.kr
Tel: +82-970-6430
ORCID:https://orcid.org/0000-0002-7055-5568

※ 이 글은 2022년도 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00751, 0.5mm급 이하 초정밀 가시·비가시 정보 표출을 위한 다차원 시각화 디지털 트윈 프레임워크 기술개발).
· Manuscript November 29, 2022; Revised January 24, 2023; Accepted January 24, 2023.

Ⅰ. 서 론

2D 이미지에서 3D 모델로의 변환은 상당히 복잡하지만 지난 몇 년 동안 많은 발전이 있었다. 이 연구 영역을 역그래픽(inverse graphics)이라고 한다. 역 그래픽 문제를 해결하기 위해서는 다른 장면 매개변수와 관련하여 해당 기능의 도함수를 계산하는 Differentiable Rendering(미분 가능한 렌더링)[1]과 5차원 데이터(x, y, z, θ, ϕ)를 입력받아

해당 시점에서 바라봤을 때 객체에서 추출되는 광선과 깊이를 추출하는 Neural radiance field(NeRF)[2] 방법이 있다. NeRF는 새로운 시점의 영상을 합성하여 3D 기하 및 외관의 장면을 압축하여 재현한다. 이후 후속 논문에 많은 영향을 주었다. 하지만 본 연구의 목적은 2D 이미지를 3D로 변환하는 기법이 재현수준을 넘어 얼마만큼 다양하고 창의적으로 생성하는지를 예술과 공학적 관점에서 실험하고 아티스트의 원본이미지와 견주어 생성 결과물을 평가하고자 한다. 이를 위해서 다양한 지오메트리와 텍스처의 표현 그리고 아티스트의 작품 제작 의도(Artist's Statement)[3]를 반영할 수 있는 의미론적 2D to 3D 변환을 탐색적으로 분석한다. 본 연구는 II장에서는 관련연구 소개와 III장은 2D to 3D 창의적 생성을 위한 제안을 하고 IV장에서는 실험 및 결과를 통해 아키텍처를 평가한다. 마지막으로 V장 결론으로 끝을 맺는다.

II. 관련연구

2D 이미지를 3D로 변환하는 기술은 3D 모델링 데이터가 필요한 일반인 또는 전문가에게 시간과 비용 절감을 위해 필요한 연구 주제이다. 하지만 전문 아티스트가 만든 높은 수준의 3D 모델과 유사하게 생성하는 기술은 쉽지 않다. 본 연구에서는 모델의 정밀도, 다양성, 창의성 등을 고려한 2D 이미지를 3D모델로 변환하는 기술과 관련된 선행연구를 소개하고, 실험을 통해 필요한 요소를 크게 4가지로 선정하였다.

첫 번째) 3D 이미지는 복원 수준을 넘어 다양한 형태로 생성되어야 한다. 2D 이미지를 3D 메쉬로 변환하는 많은 선행연구가 있으나 다양한 클래스를 보고 여러 모양으로 생성하는 모델은 상대적으로 적다. 이에 대한 연구로 StyleNeRF[4]가 있으나 그림 1(상단)에서 보는 바와 같이 카메라 포즈가 바뀌어도 배경은 그대로 유지하고 2D StyleGAN에 비해 다양성 또한 만족할 만한 수준은 아니다. Dream Fields[5]는 3D 감독 없이도 다양한 물체의 형상과 색상을 생성할 수 있는데 사전 훈련된 CLIP[6] 모델에 따라 렌더링 된 이미지가 대상 캡션과 함께 높은 점수를 받도록 많은 카메라 뷰에서 Neural Radiance Field를 최적화한다. 하지만 이 또한 최적화하는데 비용이 많이 들고 그림1(하단)처럼 명령 프롬프트에 대해 유사한 이미지를 반복적으로 생성하거나 CLIP이 가지고 있는 이미지를 사용함에 따라 그 제약을 받게 된다.

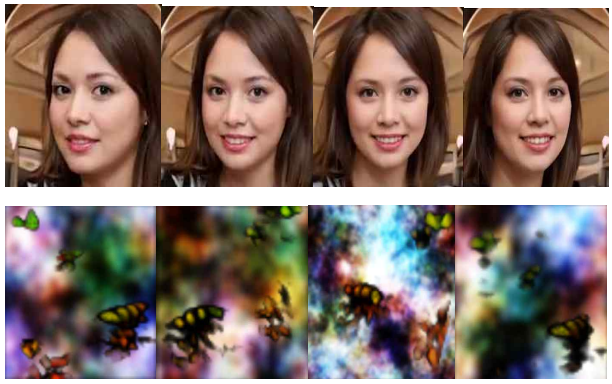

그림 1. (상단) StyleNeRF는 카메라 포즈가 바뀌어도 배경 변화가 없는 모습 (하단) Dream Fields는 자연어 쿼리로 'a world dominated by insects'를 입력으로 주어 5회 실험한 결과 유사한 이미지를 반복적으로 생성하는 모습
Fig. 1. (Top) StyleNeRF does not change the background even if the camera pose changes (bottom) Dream Fields repeatedly generates similar images after five experiments with the mouth of 'a world dominated by inserts' in a natural language query

두 번째) 2D to 3D 생성적 모델 변환에는 텍스처와 모양이 함께 변환되어야 한다. 선행연구에서는 CNN은 텍스처에 편향되어 있기 때문에 텍스처와 모양이 함께 변환되는 네트워크가 흔하지 않다고 한다[7]. 예를 들면 텍스처만 변환(Fanbo Xiang1.,2021; Lukas Hollein.,2022; Aysegul Dundar.,2022)[8,9,10]되거나 메쉬만 변환(Zhiqin Chen et al.,2021; Wang Yifan et al.,2020; Eric R. Chan et al., 2022)[11,12,13]되는 것을 볼 수 있다. 반면에 훈련 수렴을 방해하지 않으면서 조밀한 물체 표면을 학습할 수 있는 생성 복사 필드 기반의 새로운 모델인 GOF(Generative Occupancy Fields)[14] 나 기하학적 구조와 질감을 모두 나타내는 Differentiable_volumetric_rendering[15]등이 있다. 한편 AUV-Net에서는 단순히 텍스처를 표현하기보다는 UV 매핑을 통해 3D 메쉬에 연결하는 전통적인 방법이 더 바람직하다고 주장한다[16]. 이처럼 다양한 메쉬 모양과 세부적으로 디테일한 UV 매핑은 함께 적용될 필요성이 있다.

세 번째) 정밀한 3D 폴리곤 메쉬와 텍스처를 유지하면서 모바일 어플리케이션에도 최적화될 수 있게 경량화 되어야

한다. 2D 이미지를 3D로 복원 및 생성하는 많은 선행연구에서 3D 메쉬는 여전히 희미하거나 계단 현상이 나타난다. 최근 Instant-ngp[17]는 이를 개선하여 높은 해상도를 보인다. 기존의 인코딩 방법들은 하나의 태스크만 다룰 수 있었고 계산비용이 크고 속도가 낮았는데 이를 해결하기 위해 Multiresolution hash encoding을 사용해 속도를 높였다. 하지만 메쉬의 불균일성 표현으로 인한 구멍이 뚫리거나 필요 이상으로 많은 메쉬가 밀집되는 현상이 있고 3D 메쉬의 용량이 큰 단점이 있다. 이에 반해 표준 렌더링 파이프라인을 사용하여 새로운 이미지를 효율적으로 합성할 수 있는 텍스처 폴리곤을 기반으로 하는 MobileNeRF[18]는 다양한 일반 디바이스에서 실시간으로 실행할 수 있게 경량화 되어 있다.

네 번째) 창의적 2D to 3D 변환

3D 메쉬는 스타일과 형태만 다양하게 만드는 수준을 넘어 시각 예술적 관점에서 보았을 때 아티스트의 작품 제작 의도를 충분히 반영할 수 있어야 한다. 따라서 텍스트를 의미론적으로 해석(text to image)하여 원본 이미지에 충실하면서도 창의적인 3D 메쉬와 텍스처를 생성하는 것이 필요하다. 유사한 선행연구로는 다중 뷰 이미지 세트에서 알 수 없는 토폴로지, 공간적으로 변하는 재료 및 조명을 사용하여 삼각형 메쉬를 재구성하는 3D moma[19]와 3D 훈련 데이터가 필요 없이 2D 텍스트-이미지 확산 모델을 사용하여 텍스트-3D 합성하는 DreamFusion[20] 등이 있다.

III. 2D to 3D 창의적 생성을 위한 제안

2D이미지로부터 창의적인 3D 모델을 생성하기 위해 기존 소프트웨어와 딥러닝을 활용하여 그림2와 같은 방법으로 실험하였다. ①을 보면 아티스트가 만든 실제 사물을 카메라로 360° 촬영하여 약 100장~150장의 2D 이미지를 만들고 ②는 이를 3D로 복원하기 위해 3D 변환 소프트웨어인 리얼리티 캡처(Reality Capture)를 활용하였다. ③은 아티스트의 작품 제작 의도가 반영된 스타일 이미지를 메쉬에 매핑하기 위해 Diffusion CLIP[21]과 Style transfer[22]를 적용하여 다양하고 충실도 높은 텍스처를 생성하였다. ④에서는 3D 데이터의 용량을 최소화하기 위해 Blender로 버텍스의 수를 감소시키고 Unity 3D로는 텍스처의 해상도를 낮추었다. ⑤에서 SP-GAN[23]학습을 위한 메쉬와 텍스처로 학습 데이터를 구분하였다. 이후 ⑥SP-GAN을 활용하여 구체로부터 시작하여 다양한 모양의 포인트 클라우드를 생성

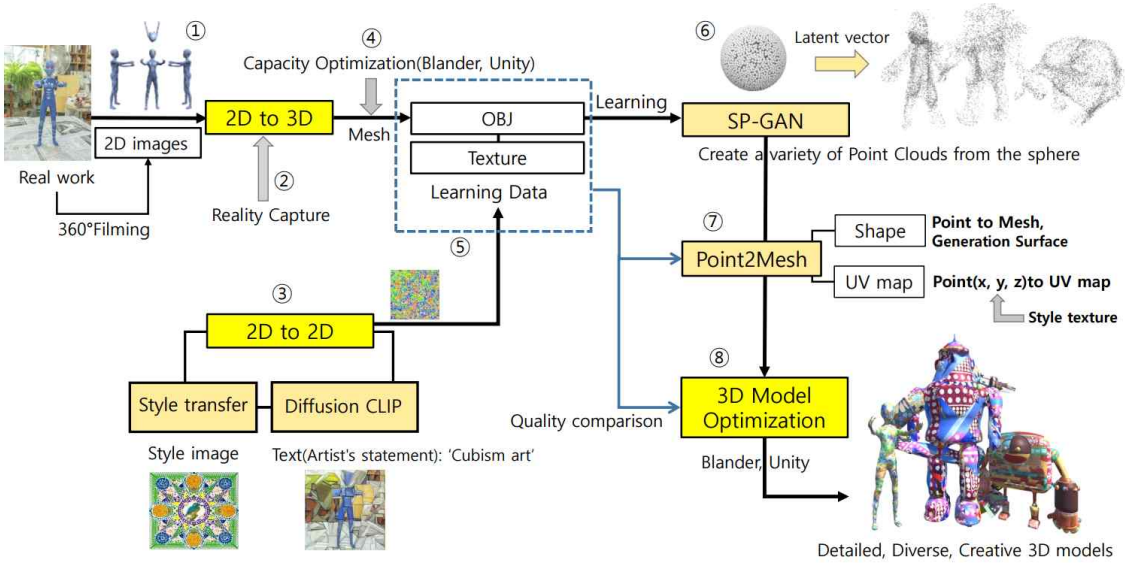

그림 2. 2D to 3D 창의적 생성을 위한 실험
Fig. 2. Experiments for 2D to 3D Creative Generation

한다. ⑦포인트 클라우드를 메쉬로 변환하기 위해서 Point2Mesh[24], Addons for blender 2.80을 활용하여 텍스처를 매핑하고 ⑧학습데이터로 사용된 메쉬와 텍스처의 품질과 유사하도록 3D 모델을 최적화하여 디테일하고 다양하면서 창의적인 3D모델을 생성하게 된다.

사용된 선행 알고리즘으로는 Diffusion CLIP, Style transfer, SP-GAN, Point2Mesh가 있으며 본 논문에서는 2D to 3D 창의적 이미지 생성을 위한 각각 알고리즘의 문제점과 이를 보완하는 실험을 통해 새로운 방법론을 제안한다. 본 논문의 기여는 다음과 같다.

- 2D이미지로부터 아티스트의 작품과 견줄만한 다양하고 창의적인 3D 모델을 생성하기 위한 방법을 제안한다. 먼저 2D이미지를 메쉬로 재현 후 학습 데이터로부터 다양하고 조밀한 포인트 클라우드를 합성한다. 최종적으로 복원된 메쉬에 아티스트스테이트먼트 텍스트에 충실한 2D 텍스처를 매핑 시킨다.
- 창작자의 스타일과 최대한 유사하면서 창의성 높은 텍스처를 생성하기 위해 스타일(StyleGAN, style transfer)만을 적용하기 보다는 작품 제작 목적이 담긴 아티스트 스테이트먼트 텍스트를 명령 프롬프트에 넣고 diffusion model의 마지막 레이어를 style transfer하여 텍스처를 생성한다.
- 창의적인 메쉬를 생성하기 위해 포인트 클라우드간 합성이후 메쉬를 추출하는 방법을 제안한다. 포인트 클라우드의 다양한 합성을 위해 기존 SP-GAN에 diffusion model과 CLIP을 적용하였고 기하학적 정확도와 미세한 포인트 클라우드를 얻기 위해 역변환 중간 레이어에 합성할 포인트 클라우드를 컨디션으로 주었다.
- Point2Mesh에서는 포인트 클라우드를 메쉬로 변환하기 위해 압축 포장한 self-prior로부터 표면 메쉬를 포인트 클라우드와 유사하게 재구성하는 방법을 사용한다. 하지만 압축 포장된 메쉬는 포인트 클라우드와 평활도 값(기하학적 상관관계 및 법선의 방향)이 다른 노이즈에 취약하여 원치 않는 모양을 만들어 내는 경우가 있다. 또한 이러한 포인트 클라우드와 유사한 압축 포장된 메쉬를 만드는 일은 쉽지 않은 일이다. 따라서 다양한 모양의 포인트 클라우드에도 강건한 최적화 기반 표면 재구성을 위해 암시적 표현과 명시적 표현을 통합하는 하이브리드 모양 표현인 SAP(Shape-As-Points)를 적용함으로써 토폴로지에 구애받지 않는 빈틈없는 매니폴드 표면을 생성한다.

IV. 실험 및 결과

그림2와 같은 소프트웨어와 딥러닝 알고리즘을 혼용한 실험을 바탕으로 이에 사용된 알고리즘의 문제점과 개선점을 파악하고 2D to 3D 창의적 생성을 위한 새로운 아키텍처를 그림3과 같이 제안한다. 실제 사물의 촬영으로부터 2D이미지를 얻고 2D to 3D 복원을 위한 Backbone으로 CLIP-NeRF[25]를 사용하여 메쉬를 만든다. 이후 SP-GAN+DDIM-CLIP을 거쳐 조밀하고 다양한 포인트 클라우드를 합성한다. 조밀한 포인트 클라우드를 얻기 위해 SAP를 적용하고 텍스처를 매핑 함으로써 창의적이고 표면이 매끄러운 메쉬를 만든다.

1. 2D to 3D 변환

2D 이미지에서 3D로의 변환은 CLIP-NeRF를 백본으로 사용한다. CLIP-NeRF는 텍스트 프롬프트를 사용하여 단일 참조 이미지(실제 사물을 촬영하여 얻은 2D이미지)로부터 보다 직관적인 방식으로 여러 3D학습데이터를 손쉽게 만드는 것이 가능하다. 실제 이미지에서 모양과 모양 코드를 유추하여 기존 데이터의 모양을 편집할 수 있다. CLIP과 StyleGAN을 결합 하여 CLIP 공간에 정의된 텍스트 조건에 따라 사전 훈련된 StyleGAN의 잠재 코드를 최적화하여 이미지를 합성한다. 따라서 실제 사물을 촬영한 2D 이미지로부터 CLIP-NeRF를 적용하여 3D 메쉬를 얻은 뒤 칼라변화, 모양 변화, 모양 제거 등의 CLIP-NeRF에서 사용한 방법에 우리는 3D 정보의 손상이 적은 3D Data Augmentation 방법을 추가하여 학습데이터의 다양성을 늘렸다.

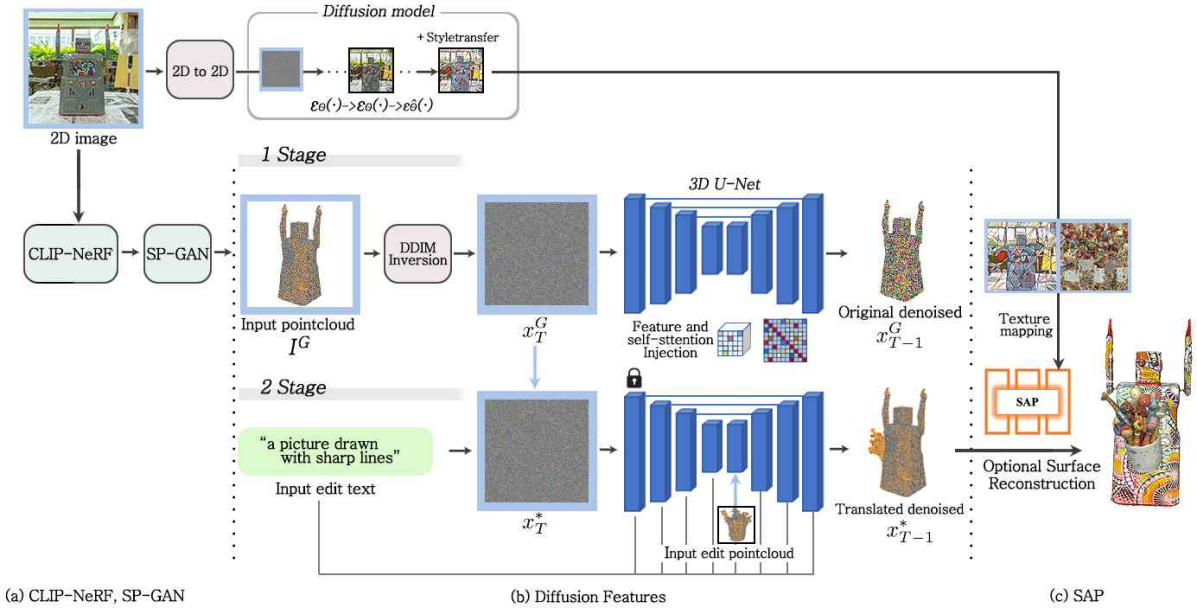

그림 3. 2D to 3D 창의적 생성을 위해 우리가 제안하는 아키텍처
Fig. 3. Our proposed architecture for 2D to 3D creative generation

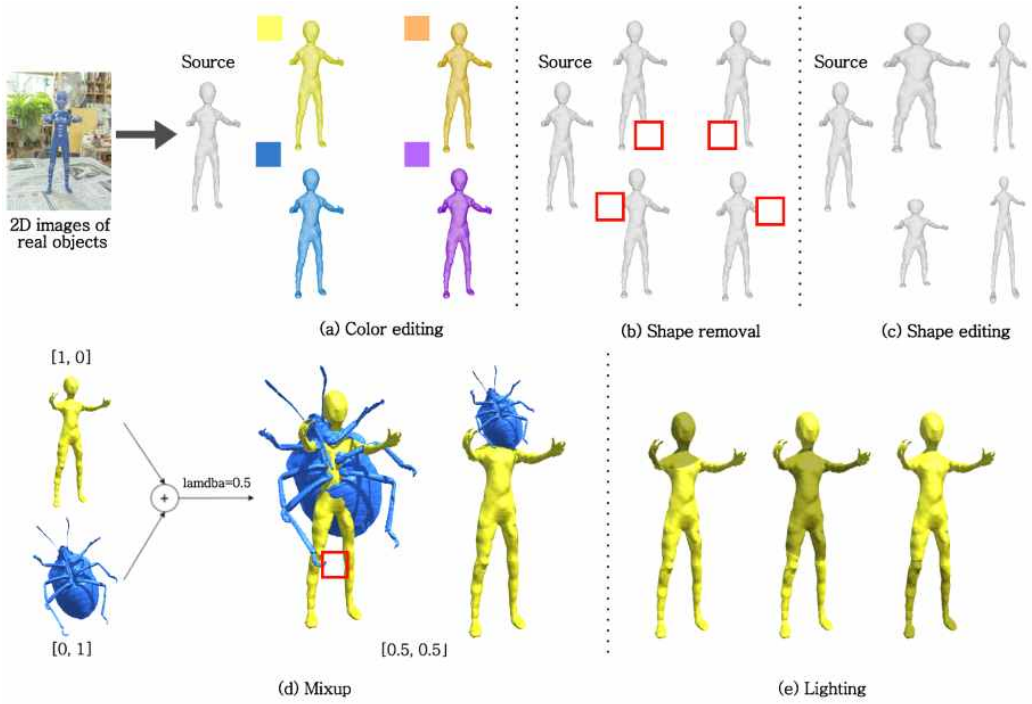

그림 4. (a),(b),(c): CLIP-NeRF를 이용한 데이터어그먼테이션, (d),(e): 우리가 추가한 데이터어그먼테이션 방법
Fig. 4. (a),(b),(c): Data augmentation using CLIP-NeRF, (d),(e): The data augmentation method we added

2. 2D to 2D 텍스처 생성

2D to 2D 다양한 이미지 생성을 위해 GAN보다 Novel detail과 다양성 면에서 성능이 좋다는 Diffusion model을 적용하였다[26]. 그림5의 (a)는 noise만 적용한 process과정이라면 (b)와 같은 우리의 방법은 CLIP Diffusion model의 DDIM Reverse단계에서 noise로 Style이미지를 넣고 복원 이미지 X0 전 마지막 레이어에 Style transfer를 적용함으로써 DiffusionCLIP만 적용한 것에 비해 아티스트 스테이트먼트(텍스트)에 가까우면서 표현력이 풍부한 텍스처가 되도록 하였다. 그림 6은 아티스트의 원본 이미지와 텍스트를 함께 diffusion model에 입력으로 넣어 생성한 이미지와

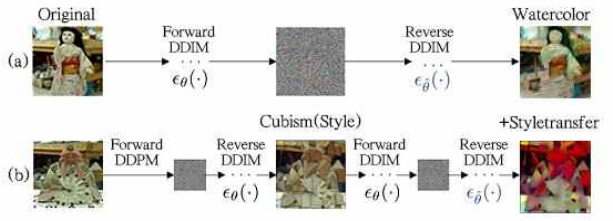

그림 5. 2D to 2D 스타일 텍스처 생성 과정
Fig. 5. Process for creating 2D to 2D style textures

Style transfer를 함께 적용한 이미지를 비교한 결과이다. 표 1을 보면 LPIPS는 두개의 이미지의 유사도를 평가하기 위해 사용되는 지표로 수치가 낮을수록 성능이 좋으나 다양한 텍스처를 만들기 위해서는 오히려 높은 것이 유리 할 수 있다.

표 1. 그림 6에 대한 성능평가
Table 1. Performance evaluation for Figure 6

Method	LPIPS ↓	PSNR ↑	SSIM ↑
Diffusion	0.562	28.14	0.244
DiffusionCLIP	0.618	28.63	0.144
DiffusionCLIP (+Style transfer)	0.710	28.01	0.260

3. 3D로부터 다양한 포인트 클라우드 생성

3D로부터 포인트 클라우드를 생성하기 위해 선행연구인 Plug-and-Play Diffusion[27]을 참조로 하였고 2D U-Net을 3D U-Net으로 바꾸었다. 레이어 중간에 합성할 포인트 클라우드를 넣었으며 Classifier free guidance의 원본이미지와 타겟 이미지의 거리를 조절하는 α를 통해 보다 정밀하고

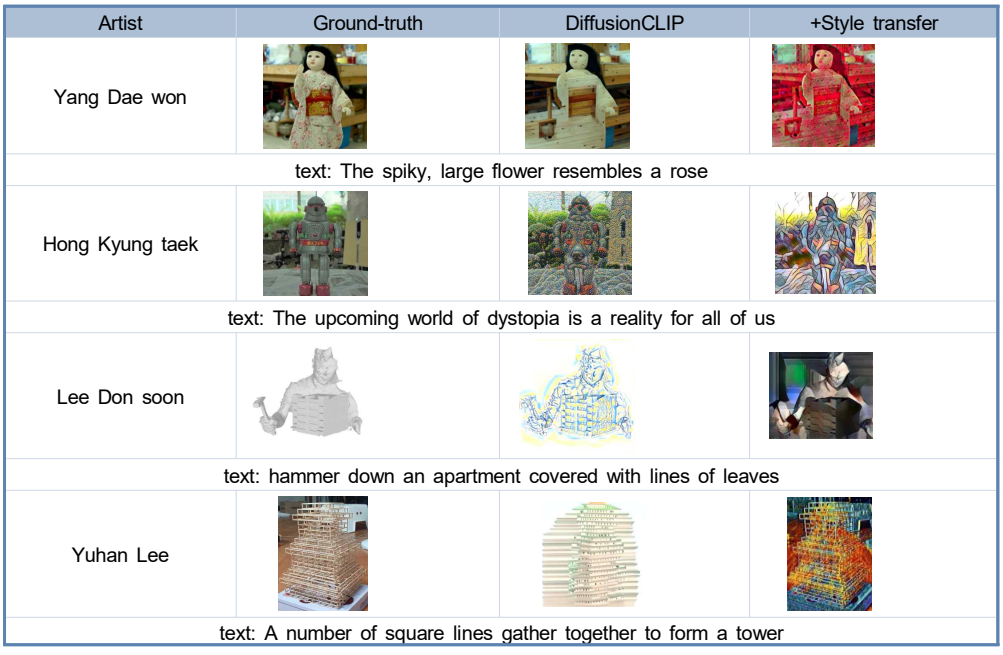

그림 6. DiffusionCLIP만 적용한 것과 Style transfer를 함께 적용한 실험 결과 이미지
Fig. 6. Image of experimental results with DiffusionCLIP only and Style transfer applied together

창의적인 포인트 클라우드를 생성하고자 하였다. 아티스트

그림 7. Global prior부터 다양한 포인트 클라우드를 생성하는 SP-GAN
Fig. 7. SP-GAN to create multiple point clouds from the Global prior

의 원본데이터를 모델링하여 학습데이터를 만든 후 SP-GAN은 그림7과 같이 구체로부터 다양하고 반추상적인 포인트 클라우드를 생성한다. 아티스트가 영감으로부터 작품을 제작하듯 SP-GAN은 적은 노이즈로 다양한 포인트 클라우드를 합성할 수 있음을 보여준다. SP-GAN의 결과는 다양한 포인트 클라우드를 생성할 수 있도록 잠재벡터를 이용한 파트별 모양 보간과 편집이 가능하나 디테일함이 부족하다. 따라서 우리가 제안하는 방법은 그림8과 같이 Diffusion model을 두개의 스테이지로 나눈 뒤 첫 번째 스테이지에서 입력 포인트 클라우드로부터 노이즈를 만들고 형태를 잘 유지시켜주기 위해 feature를 self-attention에 injection시켜 원본을 denoising한다. 두 번째 스테이지 에서는 첫 번째 스테이지에 사용된 노이즈(XT)를 가져와서 텍스트 임베딩과 합성에 사용될 포인트 클라우드를 3D U-Net 레이어에 Data Augmentation을 이용해 에디팅 시킨다. 그리고 feature와 함께 쿼리(qt)와 키(kt)를 해당 레이어에 넣어 주어 형태를 유지 시키고 다양성을 주기위해 포인트 클라우드를 컨디션으로 중간 레이어에 합성하여 Translated denoising 되게 한다. CLIP의 텍스트는 각 레이어에 컨디션으로 주어 아티스트의 원본 이미지에 의미론적으로 충실하고 다양한 포인트 클라우드의 생성이 가능하게 한다. 그림 9에서와 같이 Diffision Process 에서는 βt 에 노이즈의 정도를 조절함으로써 포인트 클라우드에 대한 역확산 과정을 특정 잠재 형태에 따라 조정되는 마르코프 연쇄로 모델링된다. 우리가 제안하는 2D to 3D 창의적 생성에는 DDPM (Denoising diffusion probabilistic models)[28]보다 더 빠르게 이미지 생성 샘플링 방법인 DDIM(Denoising Diffusion Implicit Models)[29] non-Markovian diffusion Process를 적용한다.

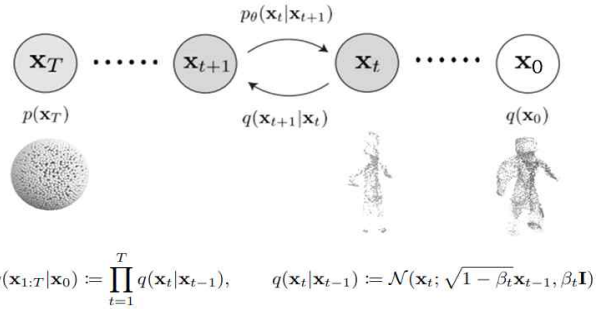

그림 9. 확산 및 생성 프로세스의 시각화. 순방향 전이 확률 $q(x_t|x_{t+1}, x_0)$이 후 $p\ominus(x_t|x_{t+1})$를 학습할 수 있다.
Fig. 9. Visualization of diffusion and generation processes. After the forward transition probability $q(x_t|x_{t+1}, x_0)$, $p\ominus(x_t|x_{t+1})$ may be learned

3.1 방법

캡션을 컨디션으로 넣어 포인트 클라우드를 다양하게 만

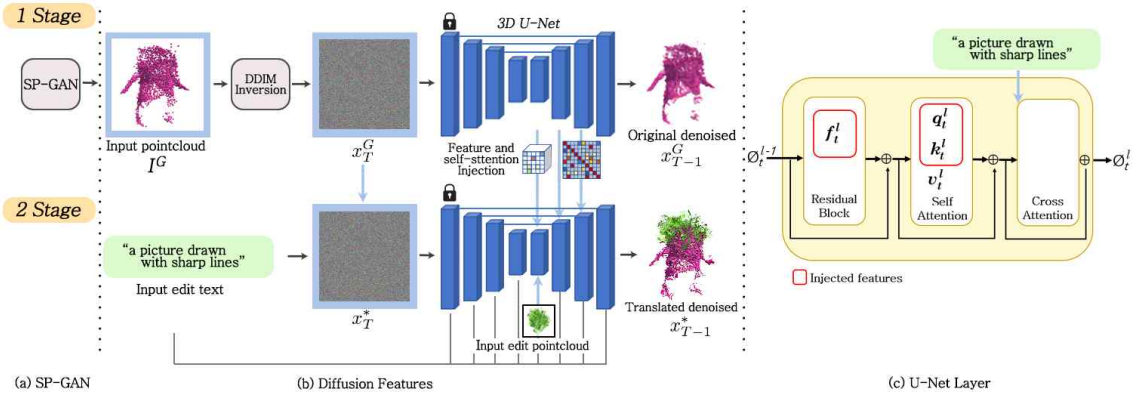

그림 8. 포인트 클라우드를 창의적으로 생성하기 위해 우리가 제안하는 방법(SP-GAN->Diffusion model+CLIP)
Fig. 8. How we propose to creatively create a point cloud(SP-GAN->Diffusion model+CLIP)

든다. 3D U-Net에 이미지를 에디팅 하고 거리 조절은 Classifier free guidance를 사용하여 α값을 조절한다.

3.1.1 Classifier free guidance

$$\epsilon = w\epsilon_\theta(x_t, P, t) + (1-w)\epsilon_\theta(x_t, \phi, t) \quad (1)$$

Classifier free guidance식은 conditional diffusion model의 아웃풋과 unconditional model의 아웃풋의 차이를 바탕으로 샘플링을 유도한다. P는 원하는 guidance를 뜻하고 ∅는 컨디션으로 주는 0 또는 null 값을 나타낸다.

(1)식에서 w>1이면 음수가 되어 원래이미지 ∅에서 멀어지는 효과가 있다. 식(2)에서는 ∅(null)를 이미지 Pn으로 교체하여 새로운 이미지(CLIP의 캡션을 임베딩 한 이미지와 텍스트)를 넣어주면 그 이미지와 멀어질수록 타켓 이미지(원하는 합성 이미지)와 가까워지게 되는 효과가 있다. α는 Pn값의 크기를 적당히 조절하여 샘플링하게 된다.

$$\tilde{\epsilon} = \alpha\epsilon_\theta(x_t, \phi, t) + (1-\alpha)\epsilon_\theta(x_t, P_n, t) \quad (2)$$

따라서 최종 식은

$$\epsilon = w\epsilon_\theta(x_t, P, t) + (1-w)\tilde{\epsilon} \quad (3)$$

으로 바뀌게 된다.

3.1.2 Self-Attention injection

Self-Attention은 Residual Block, Self-Attention, Cross Attention으로 구성된다. 입력을 제외하고 Cross-Attention 계산은 Self-Attention과 동일하다. Cross-attention은 동일한 차원의 두 개의 별도 임베딩 시퀀스를 비대칭적으로 결합하는 반면 self-attention 입력은 단일 임베딩 시퀀스이다. 시퀀스 중 하나인 Query의 역할은 feature를 그대로 입력으로 사용하여 모양을 유지하는 역할을 하고, 임베딩 벡터로 들어오는 컨디션은 Key와 Value로 들어간다.

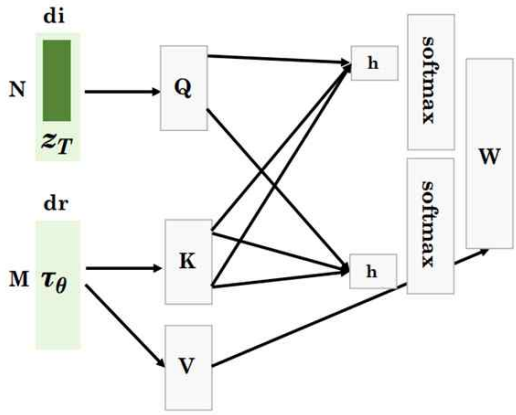

그림 11. Cross Attention 작동모습
Fig. 11. Cross Attention operation

3.1.3 포인트 클라우드 에디팅

샘플링 단계에서 합성시킬 포인트 클라우드는 컨디션으로 중간 레이어에 넣어 다양한 포인트 클라우드가 Translated denoising되게 한다. Feature and self-attention injected로 모양이 유지된 상태에서 3D U-Net 레이어 중간에 포인트 클라우드를 넣고 Shape, Mixup으로 에디팅 한다.

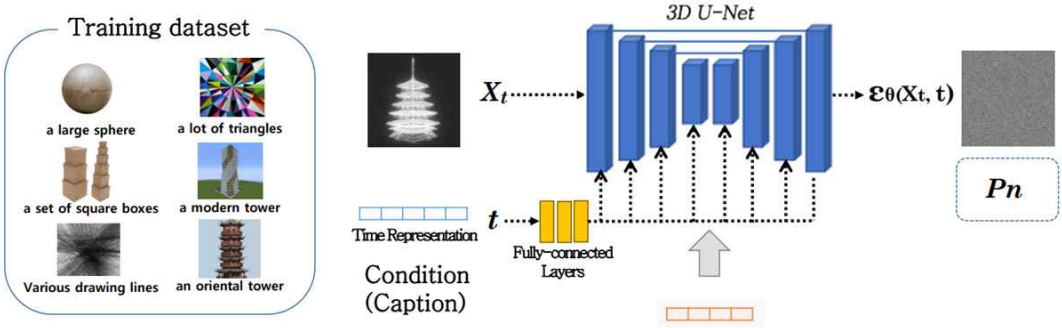

그림 10. Classifier free guidance를 적용한 이미지 유사도 거리 차이를 이용해 포인트 클라우드를 다양하게 만든다.
Fig. 10. Image similarity using Classifier free guidance makes point clouds diverse by using distance differences

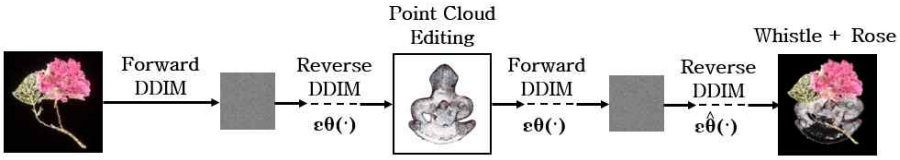

그림 12. 포인트 클라우드 에디팅을 활용한 합성
Fig. 12. Synthesis with point cloud editing

그림13은 원본과 유사하면서 창의적인 포인트 클라우드 생성을 위해 우리가 제안하는 방법과 선행 알고리즘과 비교한 실험결과이다. 표 2를 보면 우리가 제안하는 diffusion CLIP 또는 edit point cloud를 추가한 것이 좋은 성능을 보이는 것으로 나타났다. edit point cloud는 재현력 보다 창의적인 다양성을 평가하기 위한 실험으로 SP-GAN+diffusion

표 2. 3D로부터 다양한 포인트 클라우드 생성 성능평가
Table 2. Assess the performance of creating multiple point clouds from 3D

Method	LPIPS ↓	PSNR ↑	SSIM ↑	HYPE(30P/250ms)
SP-GAN	0.728	5.78	0.462	6,1,8,5
Point-E	0.629	8.07	0.520	8,3,6,10
SP-GAN (+DiffusionCLIP)OUR	0.240	15.62	0.796	7,2,3,5
SP-GAN (+DiffusionCLIP+edit point cloud)OUR	0.807	11.45	0.772	9,18,13,10

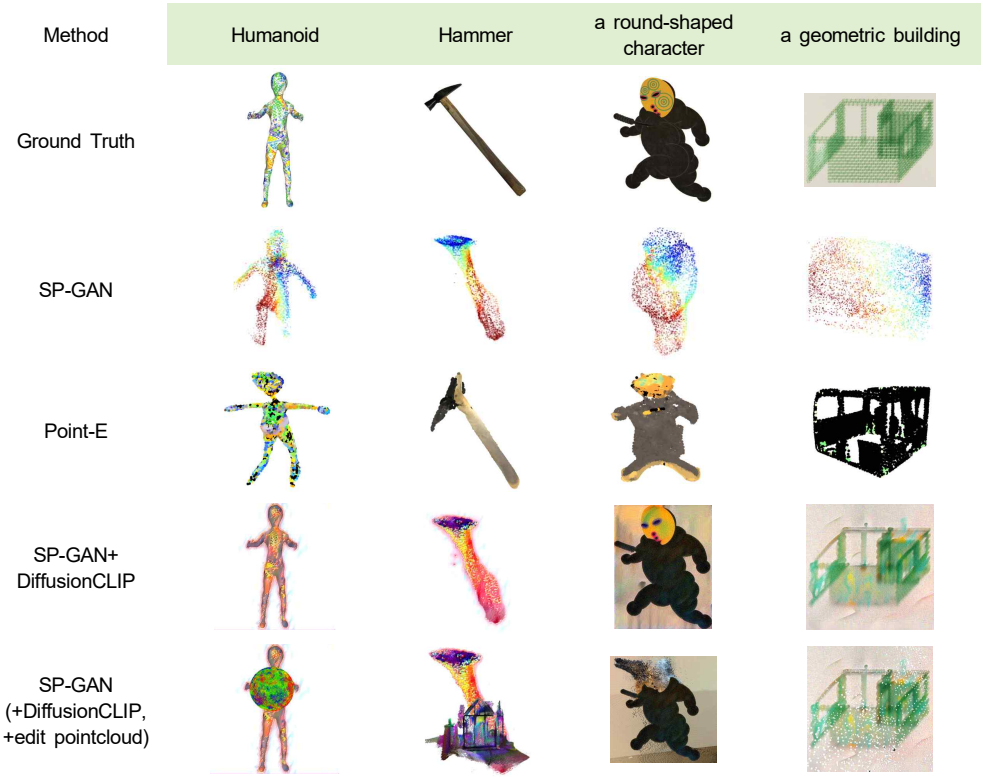

그림 13. 3D로부터 다양한 포인트 클라우드 생성 비교
Fig. 13. Comparison of creating different point clouds from 3D

CLIP에 비해 성능은 다소 떨어지나 사람이 보았을 때의 평가 방법인 Human eYe Perceptual Evaluation(HYPE)은 edit point cloud를 추가한 것이 높은 점수를 받았다. HYPE 평가방법은 HYPE-Time을 사용하여 아티스트 총 30명에게 제한된 시간 하에 시각적 인식을 측정하여 생성된 이미지가 원본과 유사하면서도 얼마나 창의적으로 표현했는지를 최소시간(250ms)안에 결정하도록 하였다. 숫자는 선호하는 사람의 수를 뜻하고 평가에 사용된 이미지는 그림13의 해당 알고리즘 좌측부터 시작하여 우측 순이다.

4. point cloud to mesh

포인트 클라우드는 Point2Mesh를 활용하여 메쉬로 변환 후 텍스처를 매핑 하였다. 그림15의 결과를 보면 포인트 클라우드를 수축 포장한 self-prior로부터 좋은 수렴을 보여준다. 하지만 포인트 클라우드와 Obj 파일의 상관구조가 기하학적으로 다르면 위에서 네 번째와 같이 노이즈가 생겨 모양을 제대로 만들지 못하는 문제점이 있다. 이것은 포인트 클라우드와 수축 포장된 메쉬와의 평활도 값(법선의 방향, 포인트 클라우드의 밀도)을 비교하여 수렴하지 않는 경우로 지정되지 않은 법선의 방향이나 저밀도 포인트 클라우드에 해당된다. 또한 메쉬 변환에는 포인트 클라우드와 압축포장된 메쉬가 함께 있어야 하는데 이러한 포인트 클라우드와 유사한 압축 포장된 메쉬를 만드는 일은 쉽지 않다. 따라서 다양한 모양의 포인트 클라우드에도 강건한 최적화 기반 표면 재구성을 위해 암시적 표현과 명시적 표현을 통합하는 하이브리드 모양 표현인 SAP(Shape-As-Points)[30]를 적용함으로써 토폴로지에 구애받지 않고 빈틈없는 매니폴드 표면을 생성하고 표현력이 풍부한 포인트 클라우드를 메쉬로 표면을 재구성할 필요가 있다.

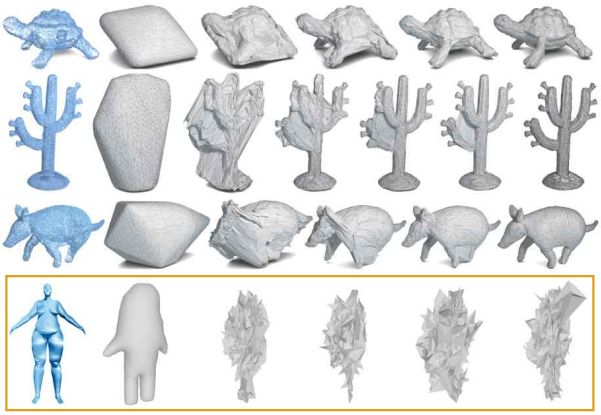

그림 15. Point2Mesh를 활용한 포인트 클라우드에서 메쉬로 변환
Fig. 15. Convert Point2 Mesh from Point Cloud to Mesh

SAP는 최적화 기반 및 학습 기반 표면 추정에 모두 사용할 수 있는 포아송 솔버(Poisson solver)가 있다.

4.1 최적화 기반 3D 재구성

최적화 기반 단일 개체 재구성을 위한 파이프라인이다. SAP를 사용하면 방향이 지정되지 않은 노이즈 포인트 클

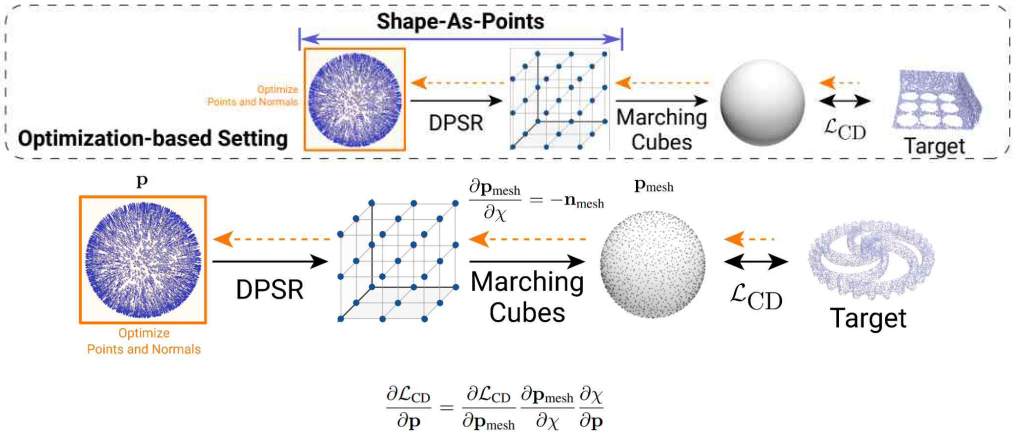

그림 14. SAP 최적화 기반 3D 재구성
Fig. 14. SAP Optimization-based Setting

라우드를 3D로 재구성을 할 수 있다. 대상 포인트 클라우드의 Chamfer loss는 최적화를 위해 정규화된 소스 포인트 클라우드로 역전파 된다.

Forward pass: 노이즈가 많은 초기 포인트 클라우드로 부터 미분 가능한 DPSR은 방향이 지정된 포인트 클라우드를 제공한다. 이러한 포아송 솔버는 점이 기본모양 내부에 있는지 여부를 나타내는 조밀한 정규 그리드(indicator grid)를 얻는다. 이후 마칭큐브를 이용하여 그리드에 일치하는 3D 구 메쉬를 출력한다. 메쉬에서 포인트를 샘플링하여 타겟과의 Chamfer Distance L_{CD}를 구할 수 있다.

Backward pass: 메시 M에서 샘플링된 모든 포인트 P_{mesh}에 대해 입력 포인트 클라우드에 대한 양방향 L2 Chamfer거리 L_{CD}를 계산한다. 모든 텀은 마칭큐브와 관련된 중간 항 P_{mesh}를 제외하고는 체인 룰을 사용하여 그라디언트를 분해할 수 있다.

4.2 학습 기반 3D 재구성

학습 기반 지표면 재구성을 위한 파이프라인이다. SAP를 사용하여 심층 신경망의 매개변수를 학습할 수 있고 신경망은 노이즈가 많은 이상 값과 지정되지 않은 포인트 클라우드의 오프셋과 법선을 모두 예측한다. 이후 보다 clean 방향의 포인트 클라우드는 DPSR(Differentiable Poisson Surface Reconstruction)로 전달되고 조건부 모델을 훈련시켜 3D로 재구성 한다. clean 방향의 포인트 클라우드를 예측하기 위해 모델을 훈련하는데, 여기서 포아송 솔버와 마칭 큐브를 사용하여 빈틈없는 메쉬를 얻는다. 한편 노이즈가 없고 조밀한 메쉬를 학습시킬 ground truth하고 이런 메쉬로 부터 표면 점과 법선을 샘플링한 뒤 PSR(Poisson Surface Re- construction)을 실행하여 정규그리드를 얻는다. 마지막으로 ground truth에 대한 L2 손실과 L_{DPSR}을 계산하고 마칭큐브에서 메쉬를 추출하게 된다.

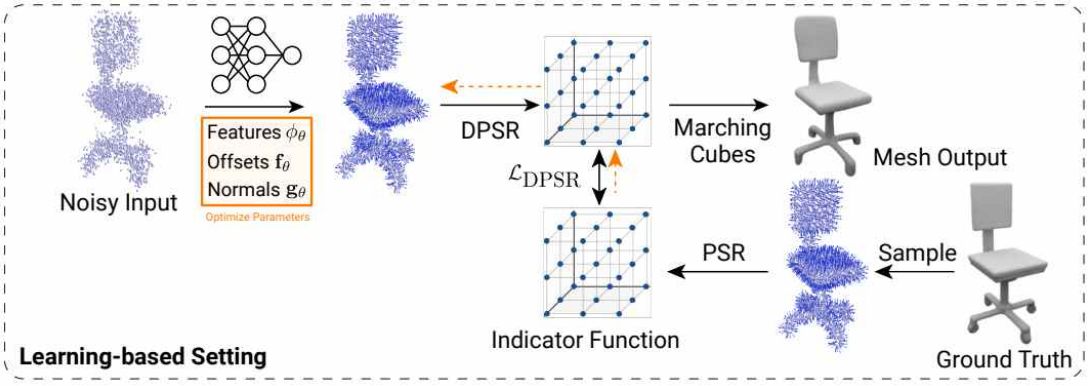

그림 16. 학습 기반 표면 재구성을 위한 파이프라인
Fig. 16. Pipeline for learning-based surface reconstruction

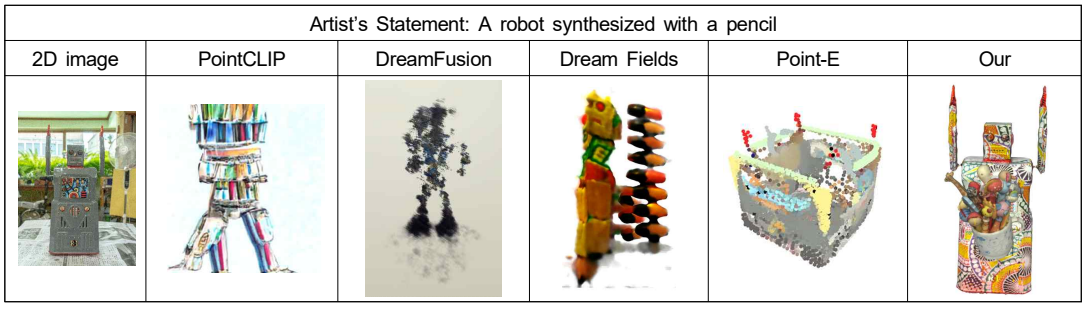

그림 17. 본 논문에서 제안하는 2D to 3D 변환 결과와 선행 알고리즘과의 이미지 비교
Fig. 17. Image comparison between 2D to 3D conversion results proposed in this paper and leading algorithms

표 3. 그림17에 대한 성능평가
Table 3. Performance evaluation for Figure 17

Method	LPIPS ↓	PSNR ↑	SSIM ↑	HYPE(30P/250ms)
PointCLIP	0.797	7.003	0.492	7
DreamFusion	0.832	9.512	0.546	2
Dream Fields	0.843	6.930	0.489	6
Point-E	0.767	7.877	0.480	4
SP-GAN+DiffusionCLIP (+edit point cloud+SAP) (Our)	0.765	10.295	0.533	11

5. 창의적 3D 모델 생성 비교 평가

그림17과 표3에서는 우리가 제안하는 2D to 3D 창의적 생성 이미지 결과와 Text to 3D 변환 선행 알고리즘이 생성한 이미지를 비교 평가 하였다. 아티스트스테이트먼트를 입력으로 넣고 PointCLIP[31], DreamFusion, Dream Fields, Point-E의 결과와 우리의 결과를 평가하는 방식이다. 표3 성능평가를 보면 우리가 제안하는 방법이 전반적으로 좋은 결과를 보이는 것을 볼 수 있다. 생성된 이미지의 화질에 대한 손실 정보 PSNR이 특히 높으며 밝기, 대조, 구조적 측면을 평가하는 SSIM에서는 DreamFusion 다음으로 좋다. 전문 아티스트 30명을 대상으로 한 HYPE 선호도에서도 높은 평가를 받았다. 평가 방법은 제한 시간 250ms안에 Ground-truth 2D이미지와 유사하면서 창의적인 3D이미지를 선택하도록 하였다.

V. 결 론

NeRF이후 2D 이미지를 3D로 변환하는 많은 연구가 진행되고 있다. 하지만 대부분 복원관련 연구가 많고 창의적으로 생성하는 분야에 대한 연구는 상대적으로 적다. 더욱이 아티스트와 견줄만한 수준의 이미지를 생성하는 방법은 어려운 태스크이다. 하지만 인공지능이 인간의 예술적 창의성에 어느 정도까지 도달 할 수 있는지를 연구한다는 것은 단순히 많은 이미지를 학습하여 생성하는 수준을 넘어서는 것이다. 즉 인간과 같이 사물을 보고 학습하여 독창적인 자기만의 작품제작의도를 바탕으로 이미지를 생성할 수 있는지에 대한 탐색적 연구는 창의적인 딥러닝 모델 발전을 위해 중요하다고 생각한다. 본 논문은 이러한 관점에서 연구하였고 최근 다양성과 정밀성 측면에서 GAN보다 성능이 좋다는 diffusion model과 CLIP을 이용하고 딥러닝과 소프트웨어를 활용한 실험을 통해 선행 알고리즘의 문제점을 파악한 뒤 2D to 3D 창의적 생성에 대한 새로운 방법을 제안하였다. 요약하면 다음과 같다.

1) 2D to 2D 텍스처 매핑

diffusion model에 Style transfer를 함께 적용하면 보다 풍부하고 다양한 텍스처 생성이 가능하다. 창작자의 스타일과 최대한 유사하면서 창의성 높은 텍스처를 생성하기 위해 스타일(StyleGAN, style transfer)만을 적용하기 보다는 작품 제작 목적이 담긴 아티스트 스테이트먼트 텍스트를 명령 프롬프트로 넣고 diffusion model의 마지막 레이어에 style transfer를 줌으로써 다양하고 풍성한 텍스처를 메쉬에 매핑 할 수 있다.

2) 창의적 2D to 3D 생성을 위한 방법

2D이미지로 부터 창의적인 3D메쉬를 곧바로 변환하는 것은 쉬운 일이 아니다. 포인트 클라우드를 거친 뒤 메쉬를 만든다고 해도 다양하고 창의적인 모양을 정밀하게 만드는 것은 어렵다. 따라서 포인트 클라우드간 합성 이후 메쉬를 추출하는 방법을 제안한다. 포인트 클라우드의 다양한 합성을 위해 기존 SP-GAN에 diffusion model과 CLIP을 적용하였고 기하학적 정확도와 미세한 포인트 클라우드를 얻기 위해 역변환 중간 레이어에 합성할 포인트 클라우드를 컨디션으로 줌으로써 창의적인 포인트 클라우드를 생성한다.

3) Point cloud to mesh

포인트 클라우드를 메쉬로 변환하기 위한 방법으로 오픈소스 라이브러리 Open3D, 메쉬를 만드는 알고리즘 Marching Cubes 등이 있다. 본 논문에서는 선행연구로 Point2Mesh를 다루었다. 하지만 압축 포장된 self-prior가

있어야 하고 포인트 클라우드와 평활도 값이 다르면 원치 않는 모양을 만들어 내는 단점이 있다. 따라서 다양한 모양의 포인트 클라우드에도 강건한 최적화 기반 표면 재구성을 위해 암시적 표현과 명시적 표현을 통합하는 하이브리드 모양 표현인 SAP(Shape-As-Points)를 적용함으로써 토폴로지에 구애받지 않는 빈틈없는 매니폴드 표면을 생성할 수 있다.

참 고 문 헌 (References)

[1] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka,Wadim Kehl and Adrien Gaidon,Differentiable Rendering, 2020, https://arxiv.org/pdf/2006.12057.pdf

[2] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, 2020, doi: https://doi.org/10.1145/3503250

[3] Artist's statement, https://en.wikipedia.org/wiki/Artist%27s_statement (Accessed November. 17, 2022)

[4] Jiatao Gu, Lingjie Liu, Peng Wang, Christian Theobalt, STYLENERF: A STYLE-BASED 3D-AWARE GENERATOR FOR HIGH-RESOLUTION IMAGE SYNTHESIS, 2021, https://arxiv.org/pdf/2110.08985.pdf

[5] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, Ben Poole, Zero-Shot Text-Guided Object Generation with Dream Fields, 2022. doi: https://doi.org/10.1109/cvpr52688.2022.00094

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Learning Transferable Visual Models From Natural Language Supervision, 2021, https://arxiv.org/pdf/2103.00020.pdf

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, Wieland Brendel, IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS, 2019, https://openreview.net/pdf?id=Bygh9j09KX

[8] Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Hao Su, NeuTex: Neural Texture Mapping for Volumetric Neural Rendering, 2021, https://arxiv.org/pdf/2103.00762.pdf

[9] Lukas Hollein, Justin Johnson, Matthias Nießner, Technical University of Munich, University of Michigan, StyleMesh: Style Transfer for Indoor 3D Scene Reconstructions, 2022, doi: https://doi.org/10.1109/cvpr52688.2022.00610

[10] Aysegul Dundar, Jun Gao, Andrew Tao, Bryan Catanzaro, Fine Detailed Texture Learning for 3D Meshes with Generative Models, 2022, https://arxiv.org/pdf/2203.09362.pdf

[11] Zhiqin Chen, Vladimir G. Kim, Matthew Fisher, Noam Aigerman, Hao Zhang, Siddhartha Chaudhuri, Simon Fraser University, Adobe Research, IIT Bombay, DECOR-GAN: 3D Shape Detailization by Conditional Refinement, 2021. doi: https://doi.org/10.1109/cvpr46437.2021.01548

[12] Wang Yifan, Noam Aigerman, Vladimir G. Kim, Siddhartha Chaudhuri, Olga Sorkine-Hornung, ETH Zurich, Adobe Research, IIT Bombay, Neural Cages for Detail-Preserving 3D Deformations, 2020. doi: https://doi.org/10.1109/cvpr42600.2020.00015

[13] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, Gordon Wetzstein, Efficient Geometry-aware 3D Generative Adversarial Networks, 2022. doi: https://doi.org/10.1109/cvpr52688.2022.01565

[14] Xudong Xu, Xingang Pan, Dahua Lin, Bo Dai, Generative Occupancy Fields for 3D Surface-Aware Image Synthesis, "Computer Vision and Pattern Recognition (cs.CV); Artificial Intelligence (cs.AI), Nov 2021. doi: https://doi.org/10.48550/arXiv.2111.00969

[15] Michael Niemeyer, Lars Mescheder, Michael Oechsle, Andreas Geiger, Max Planck Institute for Intelligent Systems, Tubingen, University of Tubingen, Amazon, Tubingen, ETAS GmbH, Bosch Group, Stuttgart, Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision, "Computer Vision and Pattern Recognition (cs.CV), jun 2020. doi: https://doi.org/10.1109/cvpr42600.2020.00356

[16] Zhiqin Chen, Kangxue Yin, Sanja Fidler, AUV-Net: Learning Aligned UV Maps for Texture Transfer and Synthesis, 2022. doi: https://doi.org/10.1109/cvpr52688.2022.00152

[17] THOMAS MÜLLER, ALEX EVANS, CHRISTOPH SCHIED, ALEXANDER KELLER, Instant Neural Graphics Primitives with a Multiresolution Hash Encoding, 2022. doi: https://doi.org/10.1145/3528223.3530127

[18] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, Andrea Tagliasacchi, MobileNeRF: Exploiting the Polygon Rasterization Pipeline for Efficient Neural Field Rendering on Mobile Architectures, 2022, https://arxiv.org/pdf/2208.00277.pdf

[19] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Muller, Sanja Fidler, Extracting Triangular 3D Models, Materials, and Lighting From Images, 2022, doi: https://doi.org/10.1109/cvpr52688.2022.00810

[20] Ben Poole, Ajay Jain, Jonathan T. Barron, Ben Mildenhall, DREAMFUSION: TEXT-TO-3D USING 2D DIFFUSION, 2022, https://arxiv.org/pdf/2209.14988.pdf

[21] Gwanghyun Kim, Taesung Kwon, Jong Chul Ye,DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation, "CVPR, Jun 2022. doi; https://doi.org/10.1109/cvpr52688.2022.00246

[22] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Image Style Transfer Using Convolutional Neural Networks,"CVPR, Jun 2016. doi: https://doi.org/10.1109/cvpr.2016.265

[23] RUIHUI LI, XIANZHI LI, KA-HEI HUI, and CHI-WING FU, SP-GAN: Sphere-Guided 3D Shape Generation and Manipulation, "ACM Transactions on Graphics, Vol.40, No.4, pp.1-12, Aug 2021.

doi: https://doi.org/10.5909/JBE.2020.25.5.776
[24] RANA HANOCKA, GAL METZER, RAJA GIRYES, DANIEL COHEN-OR, Point2Mesh: A Self-Prior for Deformable Meshes, "ACM Transactions on Graphics,Vol.39, No.4, Aug 2020.
doi: https://doi.org/10.1145/3386569.3392415
[25] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, Jing Liao, CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields, "CVPR, Mar 2022.
doi: https://doi.org/10.48550/arXiv.2112.05139
[26] Diffusion model, https://en.wikipedia.org/wiki/Diffusion_model (Accessed November. 17, 2022)
[27] Narek Tumanyan, Michal Geyer, Shai Bagon, Tali Dekel, Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation,Tue, 22 Nov 2022.
doi: https://doi.org/10.48550/arXiv.2211.12572
[28] Jonathan Ho, Ajay Jain, Pieter Abbeel, Denoising Diffusion Probabilistic Models, Wed, 16 Dec 2020.
doi: https://doi.org/10.48550/arXiv.2006.11239
[29] Jiaming Song, Chenlin Meng, Stefano Ermon, Denoising Diffusion Implicit Models, Wed, 5 Oct 2022.
doi: https://doi.org/10.48550/arXiv.2010.02502
[30] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, Andreas Geiger, Shape As Points: A Differentiable Poisson Solver, "NeurIPS, Mon 7 Jun 2021.
doi: https://doi.org/10.48550/arXiv.2106.03452
[31] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, Hongsheng Li, Shanghai AI Laboratory, Peking University, The Chinese University of Hong Kong, PointCLIP: Point Cloud Understanding by CLIP, "CVPR, Jun 2021.
doi: https://doi.org/10.1109/cvpr52688.2022.00836

저 자 소 개

조 형 래

- 2015년 : 세종대학교 회화과(서양화 전공) 학사
- 2021년 : 서울과학기술대학교 일반대학원 미디어IT공학과 석사
- 2021년 ~ 현재 : 서울과학기술대학교 일반대학원 미디어IT공학과 박사과정
- 주관심분야 : 데이터시각화, 생성모델, 실감형콘텐츠

장 일 식

- 2011년 2월 : 서울과학기술대학교 NID융합기술대학원 석사
- 2020년 3월 ~ 현재 : 서울과학기술대학교 지능형미디어연구센터 책임 연구원
- 2020년 9월 ~ 현재 : 서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학전공 박사과정
- ORCID : https://orcid.org/0000-0003-0822-9857
- 주관심분야 : 컴퓨터비전, 딥러닝

강 현 석

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

저 자 소 개

고 영 찬

- 2020년 3월 ~ 2021년 11월 : 서울과학기술대학교 지능형미디어연구센터 학부 연구원
- ORCID : https://orcid.org/0000-0002-5549-2546
- 주관심분야 : 게임 인공지능, 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 전자IT미디어공학과 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 : 컴퓨터비전, 실감미디어