# 의사 3 차원 훈련 데이터를 통한 텍스처 생성 모델의 텍스처 품질 향상

이재석, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

# Enhancing Texture Quality via Pseudo 3D Training Dataset

Jaeseok Lee, Junghyeon Seo, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

단일 이미지 텍스처(texture) 생성 과업은 3 차원 메시(mesh)와 대상 객체에 대한 한 장의 사진이 주어졌을 때 디퓨전 모델(Diffusion model)을 활용하여 해당 메시의 텍스처를 생성하는 것을 목표로 한다. 그러나 단일 이미지는 대상 객체에 대한 3 차원 표현력이 부족하기 때문에, 생성된 텍스처가 입력 이미지와 상이하다. 따라서 본 논문에서는 3 차원 인지(3D-aware) 디퓨전 모델인 Zero123 를 사용하여 훈련 데이터에 대한 3 차원 예측 이미지를 담은 3 차원 의사(pseudo) 데이터 집합을 생성하여 디퓨전 모델을 학습한 뒤 텍스처 생성 모델인 TEXTure 에 적용하도록 하여 텍스처가 원본 텍스처와 유사하도록 향상시키는 파이프라인을 제안한다. 실험 결과 의사 데이터를 사용하여 생성한 텍스처 품질이 텍스트만을 사용하였을 때나 한 장의 이미지만을 사용했을 때보다 정량적 지표, 정성적 지표 모두 객체에 대한 텍스처를 더 잘 표현함을 확인하였다.

## Ⅰ. 서 론

3 차원 컴퓨터 비전은 영화, 비디오 게임, 증강현실(Artificial Reality)과 가상 현실(Virtual Reality) 등의 산업 분야에서 많은 수요가 존재한다. 그 중에서도 텍스처를 생성하는 과업은 대상 객체에 대한 3 차원 형태를 고려하여 표면에 이미지나 패턴을 부여하는 과업이다. 따라서 3 차원 컴퓨터 비전에 대한 전문 지식을 요구하고, 작업 시간이 길기 때문에 많은 인적 자원을 소모한다. 이에 따라 인공지능 기반의 고품질의 3 차원 메시의 텍스처 자동 생성 분야의 수요는 계속적으로 증가하고 있다.

이전까지는 절대적인 3 차원 데이터 집합의 부족으로 인해 텍스처 자동 생성 분야의 발전이 더디었으나 이미지 생성 모델인 디퓨전 모델의 발전에 따라 텍스트 혹은 이미지 입력을 따르는 텍스처를 생성하는 모델들[1,2]이 등장하였다. 그러나, 입력 이미지를 따르는 텍스처 생성 과업은 객체에 대한 정보를 풍부하게 학습하기 위해 여러 장의 이미지를 필요로 한다[1]. 따라서 단일 이미지만으로 과업을 수행할 경우, 실제 외관과 차이가 발생하여 텍스처가 입력 이미지를 따르지 못할 수 있다.

따라서 본 논문에서는 3 차원 인지 디퓨전 모델인 zero123[3]를 활용하여 3 차원 의사 데이터 집합을 생성한다. 그 뒤 의사 데이터로 학습한 Dreambooth 모델을 텍스처 생성 모델인 TEXTure 에 적용하여 텍스처를 생성하는 파이프라인을 제안한다.

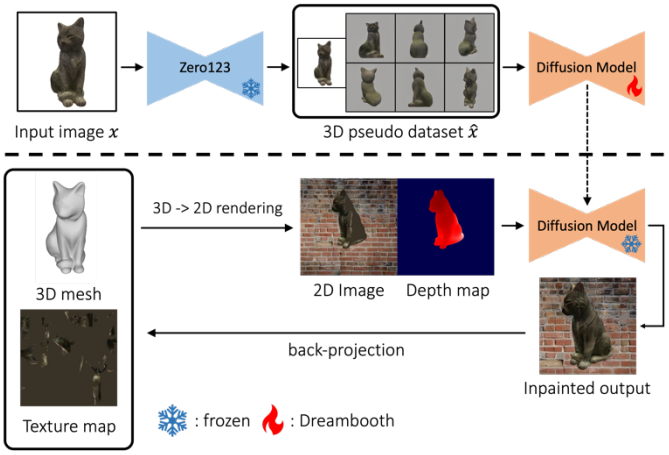


**그림 1. 제안 모델 구조.**
**(상단) 의사 데이터 집합 생성 과정**
**(하단) 메시에 대한 텍스처 생성 과정**

## Ⅱ. 본 론

디퓨전 모델(Diffusion Models)은 잡음(noise)에서 타임스텝(timestep)에 따라 점진적으로 잡음을 걷어내면서 이미지를 생성하는 생성 모델이다. 디퓨전 모델은 어텐션(attention) 연산을 통해 텍스트, 스케치, 깊이 맵 등의 다양한 조건부 이미지 생성이 가능하다.

Dreambooth[4]는 3~5 장 정도의 소규모의 훈련 데이터에 존재하는 객체(object)나 화풍(style)을 사전 학습된 디퓨전 모델에 학습시킨다. 사전 훈련된(pretrained) 디퓨전 모델에 3 장 이상의 이미지로 구성된 학습 데이터 집합에서 추출한 이미지 $x$와 텍스트 $y$를 입력하였을 때 목적 함수 $\mathcal{L}$ 은 다음과 같다.

$$\mathcal{L} = [\|\epsilon - \epsilon_\theta(x_t, t, y)\|_2^2 \qquad \text{식 (1)}$$

이때 $t$ 는 디퓨전 모델의 타임스텝(timestep), $\epsilon$ 은 디퓨전 모델이 삽입한 비 조건 잡음(unconditional noise), $\epsilon_\theta$ 은 타임스텝에 따른 노이즈를 첨가한 잡음 이미지 $x_t$, 타임스텝 $t$, 입력 텍스트 $y$ 를 조건으로 하는 조건부 잡음(conditional noise)을 의미한다. 본 논문에서는 객체에 대한 이미지를 Dreambooth 를 통해 학습시킨다.

Zero123 는 디퓨전 모델을 3 차원 객체 데이터 집합인 Objaverse[5] 데이터 집합으로 미세 조정한 3 차원 인지 생성 모델이다. Zero123 는 단일 이미지와 이미지를 어느 각도에서 보는지를 수치화한 카메라 파라미터 $\pi$ 를 입력하여 입력 이미지 $x$에 대한 해당 각도의 이미지 $x^*$을 예측한다. 이 때 $f$ 는 Zero123 모델을 의미한다. 본 논문에서는 Zero123 를 통해 입력 이미지에 대한 3 차원 의사 데이터를 생성한다.

$$x^* = f(x, \pi) \qquad \text{식 (2)}$$

TEXTure[1]는 텍스트나 이미지를 입력하여 입력을 따르는 텍스처를 생성하는 모델이다. TEXTure 모델은 다양한 카메라 각도로부터 렌더링한 깊이 맵(Depth Map)과 2 차원 이미지, 입력 텍스트를 조건으로 하여 인페인팅(inpainting) 작업을 수행한다. 그 뒤 카메라 각도와 해당 메시를 고려하여 생성한 이미지에서 새로 텍스처에 반영할 부분, 갱신할 부분, 유지할 부분으로 나누어 텍스처에

**표 1. 각 기법 간 정량적 성능 비교**

| | 텍스트 | 단일 이미지 | 의사 데이터 |
|---|---|---|---|
| LPIPS(↓) | 0.119 | 0.108 | **0.096** |

역투영(back-projection)한다. 이 과정을 반복하여 최종적으로 완성된 텍스처를 생성한다. 본 논문의 실험에서는 TEXTure 모델을 베이스라인 모델로 사용하였다.

본 논문이 제안하는 파이프라인은 다음과 같다. 우선 원본 3 차원 객체를 정면에서 렌더링하여 2 차원 이미지 $x$ 를 생성한다. 그 다음 Zero123 를 통해 $x$ 에 대한 다른 각도에서의 이미지 집합 $x_1^*, x_2^*, \cdots x_6^*$를 예측한다. 앞에서 얻은 이미지 집합에 $x$를 추가하여 3 차원 훈련 데이터 $\hat{x} = \{x, x_1^*, x_2^*, \cdots x_6^*\}$를 구성하여 Dreambooth 모델을 학습한다. 마지막으로 TEXTure 모델에 학습한 디퓨전 모델을 사용하여 3 차원 메시에 대한 텍스처를 생성한다. 전체 파이프라인에 대한 그림은 [그림 1]과 같다.

## III . 실 험 및 논 의

실험은 정량적 실험과 정성적 실험을 진행하였다. 정량적 실험에서는 객체를 설명하는 텍스트만을 사용하여 텍스처를 생성했을 때(텍스트), 원본 3 차원 객체를 정면에서 렌더링한 한 장의 이미지로 데이터 집합을 구성하여 학습한 Dreambooth 모델을 TEXTure 모델에 사용했을 때(단일 이미지), 단일 이미지에 의사 3 차원 데이터을 추가하여 학습한 Dreambooth 모델을 TEXTure 모델에 사용했을 때(의사 데이터)를 비교하였다. 텍스처가 입혀진 메시로부터 각각 360 도로 수평 회전하면서 균등하게 10 장의 이미지를 렌더링하고, 원본 텍스처 렌더링과 얼마나 시각적으로 유사한지를 측정한다. 실험 지표로는 LPIPS[6]를 사용하였다. LPIPS 는 수치가 낮을수록 지표가 좋음을 의미한다.

정성적 실험은 정량적 실험에서 사용한 기법들과 원본 텍스처를 사용한 정면, 후면 렌더링의 시각화 결과를 비교하였다. 3D 메시와 텍스처 데이터는 공공 3 차원 데이터셋을 제공하는 Poly Haven[7] 을 사용하였다.

정량적 실험 결과는 [표 1]과 같다. 가장 좋은 성능을 가진 지표는 볼드체로, 두 번째로 좋은 성능을 가진 지표는 밑줄 처리하였다. 실험 결과 의사 데이터, 단일 이미지, 텍스트 순으로 정량적 지표가 좋음을 확인할 수 있었다. 객체에 대한 캡션을 텍스트로 입력하여 텍스처를 생성하였을 경우에는 질감이나 색상 등의 객체에 대한 세부적인 정보를 디퓨전 모델에 전달할 수 없기 때문에 다른 두 기법보다 LPIPS 성능이 떨어진다. 단일 이미지의 경우, 객체에 대한 정면 정보만을 전달했기 때문에 측면 정보와 후면 정보가 포함된 3 차원 의사 데이터를 사용했을 때보다 LPIPS 성능이 떨어진다.

정성적 실험 결과는 [그림 2]와 같다. 텍스트만 사용해 텍스처를 학습한 경우에는 원본 3 차원 텍스처에 대한 색상이나 패턴, 재질이 완전히 다른 것을 확인할 수 있었다. 단일 이미지의 경우, antique ceramic vase 는 텍스트만 사용한 경우보다는 원본에 대한 패턴을 제대로 표현하지 못하는 것을 확인할 수 있다. 또한, concrete cat statue 의 경우 콘크리트 재질을 표현하지 못하는 것을 확인할 수 있다. 이는 텍스트와 단일 이미지 모두 객체에 대한 3 차원 표현력이 부족한 결과라고 추론할 수 있다.

## IV. 결 론

본 논문에서는 3 차원 인지 디퓨전 모델을 사용하여 한 장의 이미지로부터 3 차원 의사 데이터 집합을 생성하여 Dreambooth 모델을 학습하고, 학습한 Dreambooth 모델을 텍스처 생성 모델인 TEXTure 에 사용하여 텍스처를 생성하는 파이프라인을 제안한다. 실험 결과, 의사 데이터를 사용하여 생성한 텍스처가 텍스트를 사용할 때나 단일 이미지를 사용했을 때보다 객체에 대한 텍스처를 더 잘 표현함을 정량적, 정성적 실험을 통해 확인하였다.


**그림 2. 각 기법 간 시각화 결과 비교**


## ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.RS-2023-00212484, 복잡한 실제 주행환경에서 설명 가능한 움직임 예측).


## 참 고 문 헌


[1] RICHARDSON, Elad, et al. "Texture: Text-guided texturing of 3d shapes." arXiv preprint arXiv:2302.01721 (2023).

[2] CHEN, Dave Zhenyu, et al. Text2tex: Text-driven texture synthesis via diffusion models. arXiv preprint arXiv:2303.11396, 2023.

[3] LIU, Ruoshi, et al. Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 9298-9309.

[4] RUIZ, Nataniel, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 22500-22510.

[5] DEITKE, Matt, et al. Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 13142-13153.

[6] ZHANG, RICHARD, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[7] Poly Haven: The Public 3D Asset Library (https://polyhaven.com)