텍스트-이미지 임베딩 간 적응적 스케일링을 통한 카테고리 간 텍스처 전이 성능 향상

이재석, 강윤석, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Enhanced Cross-Category Texture Transfer via Adaptive Scaling of Text-Image Embeddings

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

요 약

텍스처 전이 과업은 자동으로 3 차원 메쉬의 외관을 생성하는 것을 목표로 한다. 특히, 카테고리 간 텍스처 전이는 같은 카테고리가 아니더라도 해당 이미지가 갖고 있는 재질이나 패턴을 전이할 수 있기 때문에 활용도가 크다. 그러나, 카테고리 간 텍스처 전이는 입력 메쉬와 입력 이미지의 카테고리를 모두 고려하여야 하기 때문에 두 카테고리 간 격차가 클수록 최적화가 어렵다. 따라서 본 논문에서는 IP-Adapter 기반 디퓨전 모델(Diffusion Model)에서 디퓨전 타임스텝(timestep)에 따른 텍스트-이미지 간 적응적 어텐션 스케일링 기법을 제안한다. 해당 기법을 통해 메쉬의 정보를 담은 텍스트와 텍스처 정보를 담은 이미지 모두를 고려하여 텍스처 전이가 가능함을 정량적, 정성적 실험을 통해 검증하였다.

Ⅰ. 서 론

고품질의 3 차원 컨텐츠 생성 과업은 영상 매체나 비디오 게임, 로보틱스 분야에서 필수적이다. 그 중에서도 특히 텍스처 전이(Texture Transfer)는 그래픽스 전문가의 작업 없이 자동으로 3 차원 메쉬(mesh)의 외관을 생성해 주는 과업으로, 작업 시간을 절약해 줄 수 있기 때문에 많은 연구가 진행되었다[1,2]. 특히 카테고리 간 텍스처 전이는 메쉬가 갖고 있는 고유한 특징을 고려하며 입력 이미지가 갖고 있는 재질이나 패턴을 전이하는 것을 목표로 하기 때문에 활용도가 크다[3]. 그러나, 이 과업은 메쉬와 이미지 간 카테고리 격차가 크면 최적화에 어려움을 겪을 수 있다.

따라서 본 논문에서는 IP-Adapter[4] 기반 디퓨전 모델(Diffusion Model)[5]에서 타임스텝(timestep)에 따른 텍스트-이미지 간 적응적 어텐션 스케일링 기법을 제안한다. 실험 결과 타임스텝이 클 때 텍스트 임베딩의 스케일을 크게, 타임스텝이 작을 때 이미지에 대한 임베딩의 스케일을 크게 할 때 정량적, 정성적 지표가 더 뛰어남을 확인하였다.

Ⅱ. 본 론

본 논문에서는 입력 메쉬에 대한 텍스처 공간을 표현하는 다중 퍼셉트론(MLP) θ 를 학습한다. 메쉬와 텍스처 네트워크를 통해 이미지 x_0 를 렌더링하면, 사전 학습(pretrained)한 디퓨전 모델 ϵ_ϕ는 균등분포를 따르는 무작위 타임스텝 $t\sim U(20,980)$, $t-\delta$ 에 대하여 DDIM 역전(inversion)을 수행, 노이즈가 더해진 이미지 x_t, x_s를 구한다. 본 논문에서는 δ 값을 50 으로 사용한다. 최종적으로, 디퓨전 모델은 Interval Score Matching(ISM)[6] 목적 함수를 통해 텍스처 네트워크를 학습한다.

$$\nabla_\theta \mathcal{L}_{ISM} = \mathbb{E}\left[w(t)(\epsilon_\phi(x_t,t,y)-\epsilon_\phi(x_s,t-\delta,\phi))\right] \quad \text{식 (1)}$$

이 때 $w(t)$는 타임스텝에 따라 감소하는 가중치 함수, ϕ는 빈 텍스트에 대한 임베딩, y 는 텍스트 임베딩 y_{txt} 과 이미지 임베딩 y_{img}을 합친 전체 임베딩을 의미한다.

$$y = s\cdot y_{txt} + (1-s)\cdot y_{img} \quad \text{식 (2)}$$

이 때, 텍스트 임베딩과 이미지 임베딩 간의 스케일 s를 고정 값으로 조정하면, 메쉬 정보를 담고 있는 텍스트 임베딩과 텍스처 정보를 담고 있는 이미지 임베딩 간에 불균형이 발생할 수 있다. [그림 1]에서 확인할 수 있듯이, s 값이 너무 작을 경우 텍스처 정보를 학습하지

입력 이미지 $s=0.2$ $s=0.8$ Adaptive s

그림 1. 스케일 값에 따른 텍스처 전이 결과.

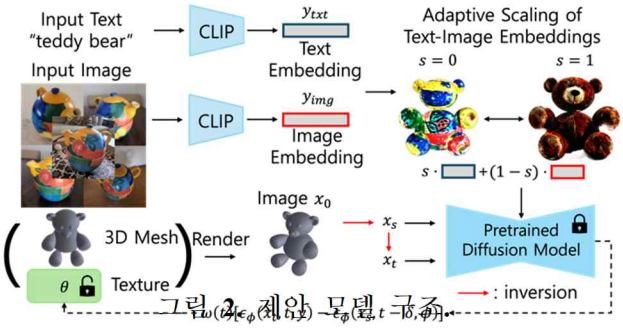

그림 2. 제안 모델 구조.

표 1. 각 모델 간 정량적 성능 비교.

	L-NeRF	TEXTure	Ours
CLIP-T(↑)	**0.370**	0.336	<u>0.362</u>
CLIP-I(↑)	<u>0.509</u>	0.446	**0.613**

못하고, s값이 너무 클 경우 메쉬 정보를 무시한다. $s = 0.4 \sim 0.6$ 일 때 대체로 좋은 결과를 얻을 수 있으나, 최적의 s 값을 찾는 것은 메쉬와 텍스처에 따라서 다를 수 있기 때문에 많은 시간을 소요할 수 있다. 따라서 본 논문에서는 디퓨전 모델이 이미지 생성 시 t 가 클 때 전체적인 부분을 생성하고, t 가 작을 때 컨텐츠를 생성한다는 기존 연구[7]에 영감을 받아 텍스트-이미지 간 적응적 어텐션 스케일링을 제안한다.

$$s = t \ / \ 1000 \quad \text{식 (3)}$$

즉, t가 클수록 텍스트 임베딩에 대한 스케일을 크게 하여 텍스처 필드 θ는 메쉬에 대한 정보를 학습한다. 반대로 t 가 작을수록 이미지 임베딩에 대한 스케일을 크게 하여 θ는 텍스처에 관한 정보를 더 학습한다. 제안 모델에 대한 전체 구조는 [그림 2]에서 확인할 수 있다.

Ⅲ. 실험 및 논의

비교 모델로는 이미지 기반 텍스처 전이가 가능한 모델인 Latent-NeRF(L-NeRF)[1] 모델과 TEXTure[2] 모델을 사용하였다. 실험 대상은 메쉬와 같은 카테고리를 가진 경우와 다른 카테고리를 가진 경우를 나눠 총 4 개의 메쉬(곰인형, 토끼, 도넛, 호박)과 5 개의 이미지로 실험을 진행하였다.

정량적 실험은 텍스트로 입력한 메쉬 정보를 잘 따르는지에 관한 CLIP-텍스트(CLIP-T)[8] 유사도, 입력 이미지에 대한 텍스처를 반영하였는지에 관한 CLIP-이미지(CLIP-I)[8] 유사도를 측정하였다. 실험 결과는 [표 1]과 같다. 실험 결과, 제안 모델이 모두 비교 모델들보다 좋은 지표를 보였다. 이는 제안 기법이 메쉬 정보를 잘 유지하면서 텍스처 전이 능력도 더 뛰어남을 추론할 수 있다.

그림 3. 각 모델 간 시각화 결과 비교.

정성적 실험은 [그림 3]과 같이 텍스처를 입힌 메쉬에 대한 시각화 결과를 비교하였다. L-NeRF 모델은 메쉬의 정보와 텍스처 정보 모두 원활하게 전달하지 못한 것을 확인할 수 있다. TEXTure 모델의 경우 같은 카테고리에 대해서는 원활하게 텍스처 전이를 수행한 반면, 카테고리 간 텍스처 전이에 대해서는 메쉬의 정보를 유지하지 못하고 텍스처 정보만 전달하였음을 확인할 수 있다. 반면 제안 모델은 메쉬 정보와 텍스처 정보 모두 비교 모델들보다 뛰어나다. 이는 텍스트-이미지 간 적응적 어텐션 스케일링 기법의 효과를 실험적으로 증명한다.

Ⅳ. 결 론

본 논문은 카테고리 간 텍스처 전이 과업을 수행하기 위해 디퓨전 타임스텝에 따른 텍스트-이미지 간 적응적 어텐션 스케일링 기법을 제안한다. 제안 기법이 메쉬 정보와 텍스처 정보 간의 균형을 유지할 수 있음을 고정 스케일링 기법과의 비교를 통해 실험적으로 확인하였다. 또한, 비교 모델들보다 카테고리 간 텍스처 전이 성능이 뛰어남을 정량적, 정성적 실험으로 검증하였다.

ACKNOWLEDGMENT

본 연구는 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2022-00167194, 미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능).

참 고 문 헌

[1] METZER, Gal, et al. Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 12663-12673.

[2] RICHARDSON, Elad, et al. Texture: Text-guided texturing of 3d shapes. In: ACM SIGGRAPH 2023 conference proceedings. 2023. p. 1-11.

[3] YEH, Yu-Ying, et al. Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 4304-4314.

[4] YE, Hu, et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

[5] ROMBACH, Robin, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 10684-10695.

[6] LIANG, Yixun, et al. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 6517-6526.

[7] CHOI, Jooyoung, et al. Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 11472-11481.

[8] RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, 2021. p. 8748-8763.