마스크를 이용한 역전 기반의 스타일 전이

최재웅, 이재구*
국민대학교 일반대학원 컴퓨터공학과

* jaekoo@kookmin.ac.kr

Inversion Based Style Transfer with Using Mask

Jaewoong Choi, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

스타일 전이(style transfer)는 목표 사진의 스타일인 특징(feature)을 추출하여 원본 사진으로 전이함으로써 원본 사진에 목표 사진의 스타일을 따르도록 하는 과업이다. 기존의 과업은 목표 사진에 대해 색상은 잘 변경하였으나, 목표 사진의 화풍이나 그림체와 같은 예술적 일관성(artistic consistency)에 대해선 고려하지 못하였다. 본 논문은 이러한 예술적 일관성을 유지하기 위해 확산 모델(diffusion model)을 이용하여 역전(inversion) 기반의 스타일 전이(style transfer)를 적용하였다. 또한 본 논문은 객체 분할 모델을 통해 얻은 마스킹(masking)으로 객체 혹은 배경에만 스타일 역전을 적용하였다. 최종적으로 원본 사진의 일부에만 스타일 역전을 적용하였으며, 추가로 이를 통해 원본 사진에 없던 객체가 생기는 문제 또한 완화되었다.

Ⅰ. 서 론

스타일 전이(style transfer)[1, 2]는 목표 사진의 특징(feature)을 추출하여 원본 사진으로 전이함으로써 원본 사진에 목표 사진의 스타일을 입히는 과업이다. 기존에는 해당 과업을 수행하기 위해 목표 사진의 특징을 추출하여 스타일을 입히는 과정을 거쳤다. 그러나 본 논문은 목표 사진의 특징을 추출하는 방법이 아닌 목표 사진을 잘 설명할 수 있는 문자 형태의 표시자(place holder)를 예측[3]하여 잠재 확산 모델(latent diffusion model)[4]에 조건(condition)으로 부여한다. 이러한 표시자는 문자 역전(textual inversion)을 통해 얻게 되며, 표시자를 이용한 스타일 전이 방법을 본 논문은 스타일 역전(style inversion)[5]이라 정의한다. 우리는 스타일 역전을 이용함으로써 기존 방법에 비해 예술적 일관성(artistic consistency)이 더 잘 보존된 결과를 확인하였다. 예술적 일관성은 단순 목표 사진에 대한 색감을 잘 유지하는 것만이 아닌, 목표 사진의 화풍이나 그림체에 대한 일관성 또한 유지되는 것을 말한다. 그러나 특정 경우에 스타일 역전을 진행하였을 때, 원본 사진에는 없던 객체가 생기는 문제가 확인됐다.

본 논문은 이러한 문제를 원본 사진의 객체 분할을 통한 마스킹(masking)[6]을 이용하여 객체 혹은 배경에만 스타일 역전을 함으로써 완화하는 방법을 제안한다. 기존 방법[5]의 경우 목표 사진에 여러 객체가 있으면, 원본 사진의 객체에 다른 객체가 추가되는 문제가 확인되었다. 본 논문은 마스킹을 이용하여 다른 객체의 추가 없이 원본 사진에서 특정 객체나 배경에 부분적으로 스타일을 바꿔주어 문제를 완화하였다. 또한 원본 사진 내 다중 객체에서도 스타일 역전이 적용되는 것을 확인하였다.

Ⅱ. 본 론

본 논문의 최종 목적은 원본 사진에서 추출한 마스크(mask)를 통해 객체 혹은 객체 외의 배경을 추출하여 해당하는 영역에 스타일(style)을 변경하는 것이다. 해당 과업을 수행하기 위해 본 논문은 크게 세 가지 모델을 사용하였다. 먼저 원본 사진에 대한 배경 혹은 객체 추출을 위해서 마스크 추출 모델[6]을 사용하였다. 다음 목표 사진을 설명하는 문자인 표시자(place holder)를 추출하기 위해 문자 역전(textual inversion) [3]을 사용하였다. 마지막으로 이

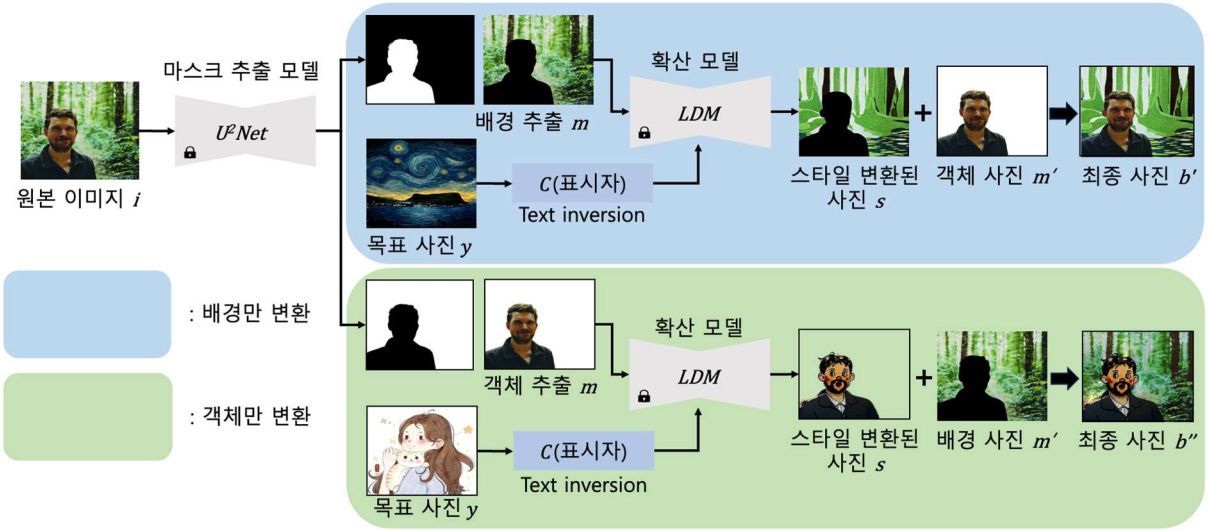

그림 1. 본 논문 제안 방법의 모델 구조

그림 2. 대상 별 스타일 변환 결과

렇게 추출된 표시자를 조건으로 넣기 위해 잠재 확산 모델(latent diffusion model)[4]을 사용하였다.

본 논문의 전체적인 모델 구조는 그림 1 을 통해 확인할 수 있다. 먼저 객체 혹은 배경 m 을 추출하기 위해 본 논문은 U^2Net [6]을 사용하였다. U^2Net 은 객체 분할에 사용되는 두 단계의 U 구조 모델이며, 원본 사진 i 를 입력을 받을 때, 배경 혹은 객체 m 을 출력으로 가진다.

$$m = U^2Net(i) \quad (1)$$

다음 스타일을 가져오려는 사진인 목표 사진 y 에 대해 잘 설명하는 표시자 C 를 가져오기 위해 문자 역전 모델을 사용하였다. 이러한 문자 역전으로 얻은 임베딩(embedding)은 목표사진에 대해 잘 설명하는 문자역할을 하며, 확산 모델에서 조건(condition)으로 사용될 수 있다.

$$C = argmin\ E_{z,x,y,t}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, c_\theta(y)\right)\right\|_2^2\right] \quad (2)$$

여기서 y 는 그림 1 의 목표 사진이며, C는 구하고자 하는 표시자이며, z_t는 시점 t의 잠재 코드(latent code), ε 은 순방향 확산 과정에 생성된 노이즈(noise)이다. $\varepsilon_\theta\left(z_t, t, c_\theta(y)\right)$는 역방향 확산 과정에서 t 시점에 예측한 노이즈이며, $c_\theta(y)$ 는 학습가능한 벡터(vector)이다. 최종적으로 ε 와 ε_θ 의 값의 차이를 최소화함으로써 표시자 C를 구하게 된다.

마지막으로 구해진 표시자 C 를 잠재 확산 모델에 조건으로 넣어준다. 잠재 확산 모델은, 확산 기반의 생성 모델로, 조건을 입력으로 받아 조건에 적합한 사진을 생성한다. 단, 기존 잠재 확산 모델은 추론 시 무작위 노이즈를 사용하나, 본문은 추출된 배경 혹은 객체에 조건을 주기 위해 추출된 배경 혹은 객체에 노이즈를 추가하여 사용한다.

$$s = LDM(z', C) \quad (3)$$

여기서 LDM 은 잠재 확산 모델(latent diffusion model)이며, z'은 배경 혹은 객체인 m에 노이즈가 추가

그림 3. 다중 대상에서의 스타일 변환 결과

된 잠재 코드이다. 이로써 목표 사진의 스타일이 원본 사진에 전이된 결과인 s가 생성된다.

마지막으로 전이된 결과인 s 와 배경 혹은 객체 사진 m' 을 더해주어 최종 사진인 b' 혹은 b'' 을 생성한다. 이를 통해, 원본 사진에 대해 객체 혹은 배경에 대해서만 스타일 역전을 진행하였다.

Ⅲ. 실험

본 논문은 배경과 사람을 잘 드러내는 데이터 집합을 구하기 위해 잠재 확산 모델[4]에서 사진을 직접 생성하여 사용하였다.

그림 2 는 객체와 배경에 각각 스타일 역전을 한 결과이다. 다음 결과를 통해 객체에만 스타일 역전을 적용하였을 때, 원본 사진에는 없던 객체가 생성되는 문제가 해소된 것을 확인하였다. 그림 3 에서는 다중 객체에서도 객체 분할과 스타일 역전이 잘 적용되는 것을 시각화 하였다.

평가지표는 사용자 선호도 조사를 사용하였다. 사용자 선호도 조사는 12 명을 대상으로 진행하였으며, 객체가 자연스럽게 분할되어 배경 혹은 객체에만 스타일이 잘 입혀졌는지를 평가 점수를 통해 정성적으로 평가했다. 평가 점수는 만족도에 따라 1 점~5 점 사이로 16 명을 대상으로 책정하여 결과에 대한 점수를 산정했다. 사용자 선호도 점수는 3.56 점으로 선호 정도가 보통 이상임을 확인하였다.

Ⅳ. 결 론

본 논문은 기존의 마스킹 모델과 스타일 역전 모델을 결합하여 최종적으로 원본 사진에 대해 객체 혹은 배경에 대해서만 스타일 역전을 진행하였다. 이러한 과정을 진행하였을 때, 우리는 원본 사진에 추가적인 객체가 생성되는 문제를 완화하였다. 또한 스타일 역전 통해 예술적 일관성을 잘 보존하는 결과를 생성하였으며, 객체 혹은 배경의 분할 또한 자연스럽게 되는 것을 확인하였다.

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] StyTr2: Image style transfer with transformers. In. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11326-11336, 2022. 2, 3, 6, 7, 8

[2] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Do- main enhanced arbitrary image style transfer via contrastive learning. In *ACM*

SIGGRAPH 2022 Conference Proceedings, pages 12:1–12:8, New York, NY, USA, 2022. Association for Computing Machinery.

[3] Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." *arXiv preprint arXiv:2208.01618* (2022).

[4] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[5] Zhang, Yuxin, et al. "Inversion-based style transfer with diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[6] Qin, Xuebin, et al. "U2-Net: Going deeper with nested U-structure for salient object detection." *Pattern recognition* 106 (2020): 107404