이미지 잔물결 제거로 자연스러운 이미지 속성 편집이 가능한 StyleFusion 기반의 High-Fidelity GAN Inversion

StyleFusion-based High-Fidelity GAN Inversion for Natural Image Attribute Editing by Removing Image Ripples

요 약

최근 GAN inversion 을 기반으로 사진 합성 기술이 연구되고 있다. 이미지를 고화질로 재구성 및 편집하려면 이미지 세부 정보까지 보존해야 하는데, 재구성과 편집 사이에는 trade-off 문제가 있다. 최신의 GAN Inversion 프레임워크인 HFGI 는 재구성과 편집에서 균형을 맞춰 이 문제를 극복하였지만, 시점 변경이 큰 이미지를 편집할 때 잔상이 남는 치명적인 한계가 있다. 본 연구에서는 시점 변경이 큰 이미지를 편집할 때도 충분히 개선될 수 있도록 HFGI 에 StyleFusion 을 융합한 새로운 모델 HFGI-SF 을 제안한다.

1. 서 론

최근 GAN 기술이 발전함에 따라 이미지를 편집할 수 있는 방법들이 많이 연구되고 있다. 이미지 편집할 때는 원하는 속성(예: 표정, 나이)을 수정하되, 다른 세부 사항들은 유지될 수 있어야 한다. 예를 들어 StyleGAN 은 여러 이미지 생성모델 중에서 이미지를 재구성하고 편집하는 데에 있어서 왜곡이 낮고 품질 또한 뛰어나다 [1]. 하지만 현존하는 GAN inversion 모델들은 이미지 편집에 고질적인 한계가 존재한다. 인코더 기반의 GAN Inversion 기술은 실제 이미지를 저차원의 latent code 로 압축하게 되면 필연적으로 정보 손실이 일어나게 되는데, 손실되는 정보는 주로 이미지의 세부 정보이다. 이미지를 고화질로 재구성 및 편집을 하려면 이러한 세부 정보까지 모두 보존해야 할 필요가 있다.

현존하는 GAN inversion 모델 중에서 High-Fidelity Gan Inversion 은 재구성과 편집이 모두 개선이 잘 이루어졌다[2]. 하지만 극단적인 misalignment 인 이미지 사례를 처리할 때 원본 이미지의 잔상이 남는 치명적인 문제가 발생하는 것을 발견하였으며, 그림 1 에서 확인할 수 있다. 따라서 HFGI 프레임워크에서 시점 변경이 큰 이미지를 편집할 때, 잔상(잔물결)이 없이 이미지 편집이 가능하도록 개선된 모델을 제안한다.

본 논문의 구성은 다음과 같다. 2 장에서는 GAN 과 GAN Inversion, Rate-Distortion-Edit Trade-Offs 그리고 HFGI 프레임워크에 대해 설명을 하고, 3 장에서는 제안된 모델의 구조를 제시한다. 4 장에서는 실험을 통해 기존의 HFGI 과 개선된 모델의 성능을 비교한다. 마지막으로 5 장에서는 결론을 맺는다.

2. 관련 연구

2.1. GAN

GAN (Generative Adversarial Networks) 은 생성자(Generator)와 구분자(Discriminator) 두 네트워크를 적대적(Adversarial)으로 학습시키는 비지도 학습 기반의 생성모델(Unsupervised generative model)이다. GAN 은 생성자가 만든 가짜 데이터가 진짜 데이터와 비슷하여 판별자가 진위를 판별하지 못할 때까지 알고리즘을 개선하는 방식으로 학습을 진행한다. 이처럼 GAN 으로 학습하는 생성자는 진짜 같은 가짜 데이터를 만들어내기 때문에, 유명 화가의 화풍을 입힌 이미지나 음성 변조 파일, 영상 등 다양한 콘텐츠 분야에서 활용되고 있다.

2.1.1. GAN Inversion

GAN Inversion 은 GAN 과는 반대로, 주어진 실제 이미지를 재구성할 수 있는 가장 적절한 latent code 를 찾아내는 기술이다. 기존의 GAN Inversion 접근법으로는 세 가지로 분류가 가능하다. (1) 최적화 기반 (2) 인코더 기반 (3) 하이브리드. 최적화 기반의 GAN Inversion 기술로는 I2S, PTI 등이 있으며 이미지 별로 최적화가 이루어지기 때문에 재구성의 품질이 뛰어나지만 추론하는 과정에서 시간이 많이 소모되고 편집 능력이 떨어진다 [3,4]. 반면에 인코더 기반은 편집 능력이 좋고 빠른 추론이 가능하지만 실제 이미지가 latent code 로 압축되는 과정에서 필연적으로 정보 손실이 발생하여 재구성의 품질이 좋지 않으며, 예시로 pSp 와 e4e 등이 있다 [5,6].

그림 1. 원본 이미지와 시점 변경한 편집 이미지

2.1.2. Rate-Distortion-Edit Trade-Offs

GAN Inversion 기술에서 Information bottleneck 이론에 따르면, 깊은 압축으로 인해 손실되는 정보는 주로 이미지 세부 정보(high-frequency 패턴)이다. Low-rate latent code 은 차원이 낮으므로 일부 정보가 불가피하게 손실되고, High-dimension 으로 차원을 올리는 방법은 reconstruction 품질이 좋아지지만, 과적합이 되기 쉬워져서 편집성이 떨어진다. 이처럼 Reconstruction 과 Editability 는 trade-off 관계에 있다. 따라서 이미지 생성 및 편집하는 기술은 editability 의 성능을 손상시키지 않으면서 reconstruction 의 성능(fidelity)을 올릴 수 있도록 균형이 잘 맞는 섬세한 시스템의 설계가 필요하다.

2.2. High-Fidelity GAN Inversion

High-Fidelity GAN Inversion 은 인코더 기반으로, DCI 방법과 ADA 모듈을 제안하여 이미지의 세부 정보가 잘 보존된 속성 편집을 가능하게 하였다.

2.2.1. Distortion Consultation Inversion (DCI)

low-rate latent vector 이 놓쳐버린 high-frequency 한 이미지 세부 정보를 갖는 “distortion map”을 활용한다. Distortion map 은 무시된 이미지 세부 정보를 다시 가져와 해당 논문에서 새로 고안된 consultation 인코드를 통해 high-rate latent map 에 투영되고 low-rate latent vector 과 융합을 한다. 네트워크는 generation 을 위한 참조로 이미지 세부 정보를 명시적으로 consult 하게 되며 기존의 기본 인코더가 보완이 된다.

2.1.2. Adaptive Distortion Alignment (ADA)

이미지를 속성 편집할 때 low-rate latent code W 는 다음과 같이 특정한 semantic 방향에 따라 이동한다: $W^{edit} = W + \alpha N^{edit}$ 이 경우 general 하게 작동하지만 distortion map 에서 inversion 과는 다르게 편집된 이미지는 misalignment 가 발생하는 문제가 있다. ADA 모듈은 이러한 misalign 을 잡아줄 수 있다. ADA 는 encoder-decoder 구조이며, self-supervised learning 을 한다. Ground Truth 는 기존의 distortion map 이고, 랜덤한 augment 를 distortion map 에 적용한 것과 GT 의 차이를 L1 loss 로 적용하여 최적화를 한다. alignment loss 는 다음과 같다: $L_{align} = \|\hat{\Delta} - \Delta\|_1$

3. 방법

HFGI 은 사람 얼굴 영역에서 InterfaceGAN 을 채택하여 이미지를 편집하였다[7]. 편집가능한 속성 중에서 pose 를 극단적으로 변경을 하면, 그림 1 과 같이 원본 이미지의 잔상이 남는 치명적인 문제가 발생한다. 이를 해결하기 위해 StyleFusion 의 구조를 채택하여 기존 HFGI 와 융합하는 새로운 모델 HFGI-SF 을 제안한다[8]. 제안하는 모델의 구조는 그림 2 에 묘사되어 있다.

3.1. 모델 구조

새롭게 제안하는 HFGI-SF 모델의 구조는 그림 2 에 묘사되어 있으며, 다음과 같다: 사전훈련이 된 e4e 기반의 인코더 E_0 와 StyleGAN2 기반의 생성기 G_0 에 의해 반전된 첫 이미지 $\hat{X}_0$ 가 생성된다. 생성된 이미지는 관심이 있는 의미론적 영역인 ‘배경’과 ‘인물’ 총 2 가지로 공간적 분리를 한다. 저차원의 잠재 코드 w 는 이미지 $\hat{X}_0$ 의 세부 정보를 놓치게 되며, 후에 DCI 를 통해 해당 정보를 융합할 수 있도록 distortion map $\tilde{\Delta}$ 을 계산한다. 특정 방향으로 속성 편집이 된 이미지 $\hat{X}_0^{edit}$ 는 정렬이 왜곡될 수 있기 때문에, $\tilde{\Delta}$ 과 함께 입력으로 들어가서 ADA 모듈을 통해서 재정렬을 하도록 훈련을 한다. 출력으로 나온 $\hat{\Delta}^{edit}$ 은 pSp 기반의 consultation 인코더 E_C 의 입력으로 들어가 DCI 를 통해 최종으로 편집된 이미지 $\hat{X}^{edit}$ 가 생성된다. $\hat{X}^{edit}$ 는 $\hat{X}_0$ 과 마찬가지로 관심이 있는 의미론적 영역인 배경과 인물로 분할을 한다. 이때 $\hat{X}_0$ 의 ‘배경’과 $\hat{X}^{edit}$ 의 ‘인물’은 FusionNet 에 들어가 Latent Blender 에 의해 잠재 코드가 정렬되고, 두 잠재 코드의 의미 영역을 융합하는 방법을 학습하여 최종적으로 배경과 인물이 융합된 이미지가 생성이 된다.

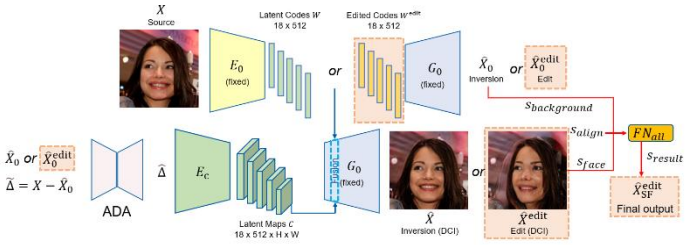

그림 2. HFGI-SF 구조

4. 연구 및 결과

4.1. 데이터

사람 얼굴 영역에서 실험을 진행하며, 훈련에는 FFHQ 데이터셋을 사용하고 모든 이미지는 1024x1024 크기이다[9]. 테스트를 진행할 때는 CelebA-HQ 중 테스트 데이터셋을 사용한다 [10].

5. 결론

최근 GAN inversion 기술은 이미지의 세부 정보를 보존하도록 발전하고 있지만, 아직까지도 한계가 존재한다. 우리는 고화질 이미지 속성 편집을 가능하게 하는 high-fidelity GAN inversion 프레임워크에서 시점 변경이 큰 이미지를 속성 편집할 때 원본 이미지의 잔상(잔물결)이 남는 치명적인 문제점을 발견하였다. 이를 해결하기 위해 HFGI 와 StyleFusion 을 융합하여 배경과 인물을 따로 분리하고 마지막에 조합하여 이미지의 잔물결이 제거될 수 있도록 하는 방법을 제안한다. 제안된 모델은 재구성 및 편집에서 이미지별 세부 정보 보존이 잘되어 있으며, 시점 변경이 큰 이미지 편집에서도 잔물결이 없는 명확한 개선을 보여준다.

참 고 문 헌

[1] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[2] Wang, Tengfei, et al. "High-fidelity gan inversion for image attribute editing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[4] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744, 2021.

[5] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[6] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG), 40(4):1-14, 2021.

[7] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[8] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel CohenOr. Stylefusion: A generative model for disentangling spatial segments. arXiv preprint arXiv:2107.07437, 2021.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. The International Conference on Learning Representations (ICLR), 2018.