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

임에딘

경희대학교 컴퓨터공학과

vlfdus0401@khu.ac.kr

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

Edin Lim

Computer Science and Engineering, Kyung Hee University

요 약

최근 GAN inversion을 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 HFGI는 재구성과 편집에서 균형을 맞춰 이 문제를 극복하였지만, 시점 변경이 큰 이미지를 편집할 때 잔상이 남는 치명적인 한계가 있다. 본 연구에서는 시점 변경이 큰 이미지에서 개선하기 위해 hfgi에 graphonomy를 합쳐서 배경과 인물을 분리하여 인물 부분에서는 잔물결 부분을 masking을 하고, 배경 부분에서는 lama inpainting기법을 통해 합성하여 결과적으로 이미지의 잔물결을 제거할 수 있도록 한다.

1. 서론

1.1. 연구배경

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 StyleGAN은 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 code로 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 Inversion은 재구성과 편집이 모두 개선이 잘 이루어졌다[2]. 하지만 극단적인 misalignment인 이미지 사례를 처리할 때 원본 이미지의 잔상이 남는 치명적인 문제가 발생하는 것을 발견하였으며, 그림 1에서 확인할 수 있다. 따라서 HFGI 프레임워크에서 시점 변경이 큰 이미지를 편집할 때, 잔상(잔물결)이 없이 이미지 편집이 가능하도록 개선된 모델을 제안한다.

1.2. 연구목표

인코더 기반의 large-rate GAN inversion system으로 이뤄진 HFGI 프레임워크에서의 한계점이었던 extreme misalignment의 이미지들이 편집할 때 충분히 개선될 수 있도록 기술을 연구하는 것이 본 연구목표이다.

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. GAN은 생성자가 만든 가짜 데이터가 진짜 데이터와 비슷하여 판별자가 진위를 판별하지 못할 때까지 알고리즘을 개선하는 방식으로 학습을 진행한다. 이처럼 GAN으로 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 Inversion은 GAN과는 반대로, 주어진 실제 이미지를 재구성할 수 있는 가장 적절한 latent code를 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 code로 압축되는 과정에서 필연적으로 정보 손실이 발생하여 재구성의 품질이 좋지 않으며, 예시로 pSp와 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 code은 차원이 낮으므로 일부 정보가 불가피하게 손실되고, High-dimension으로 차원을 올리는 방법은 reconstruction 품질이 좋아지지만, 과적합이 되기 쉬워져서 편집성이 떨어진다. 이처럼 Reconstruction과 Editability는 trade-off 관계에 있다. 따라서 이미지 생성 및 편집하는 기술은 editability의 성능을 손상시키지 않으면서 reconstruction의 성능(fidelity)을 올릴 수 있도록 균형이 잘 맞는 섬세한 시스템의 설계가 필요하다.

2.2. High-Fidelity GAN Inversion

High-Fidelity GAN Inversion은 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 vector이 놓쳐버린 high-frequency한 이미지 세부 정보를 갖는 "distortion map"을 활용한다. Distortion map은 무시된 이미지 세부 정보를 다시 가져와 해당 논문에서 새로 고안된 consultation 인코드를 통해 high-rate latent map에 투영되고 low-rate latent vector과 융합을 한다. 네트워크는 generation을 위한 참조로 이미지 세부 정보를 명시적으로 consult 하게 되며 기존의 기본 인코더가 보완이 된다.

2.2.2 Adaptive Distortion Alignment (ADA)

이미지를 속성 편집할 때 low-rate latent code W는 다음과 같이 특정한 semantic 방향에 따라 이동한다: $W^{edit} = W + \alpha N^{edit}$ 경우 general하게 작동하지만 distortion map에서 inversion과는 다르게 편집된 이미지는 misalignment가 발생하는 문제가 있다. ADA 모듈은 이러한 misalign을 잡아줄 수 있다. ADA는 encoder-decoder 구조이며, self-supervised learning을 한다. Ground Truth는 기존의 distortion map이고, 랜덤한 augment를 distortion map에 적용한 것과 GT의 차이를 L1 loss로 적용하여 최적화를 한다. alignment loss는 다음과 같다: $L_{align} = \|\hat{\Delta} - \Delta\|_1$

3. 프로젝트 내용

3.1. 시나리오

HFGI은 사람 얼굴 영역에서 InterfaceGAN을 채택하여 이미지를 편집하였다[7]. 편집가능한 속성 중에서 pose를 극단적으로 변경을 하면, 그림 1과 같이 원본 이미지의 잔상이 남는 치명적인 문제가 발생한다. 이를 해결하기 위해 hfgi에 graphonomy[11]를 합쳐서 배경과 인물을 분리하여 인물 부분에서는 잔물결 부분을 masking을 하고, 배경 부분에서는 lama inpainting[12]기법을 통해 합성하여 결과적으로 이미지의 잔물결을 제거할 수 있도록 한다.

3.2. 요구사항

인코더 기반의 large-rate GAN inversion system으로 이뤄진 HFGI 프레임워크에서 한계점이었던 extreme misalignment의 이미지들 또한 편집할 때 충분히 개선이 되어야 한다. 본 연구에서는 이미지를 편집하는 속성들 중에서도 "pose" 부분이 편집될 때 생기는 '잔물결' 현상을 최대한 없애고 자연스러운 이미지 편집을 하는 것이 요구된다. 이를 위해 graphonomy를 hfgi에 추가하여, 편집하고자 하는 원본 이미지를 foreground(인물) 부분과 background(배경) 부분으로 나눈다. 그 다음, 인물 부분만을 hfgi editing (interfaceGAN)을 통해 pose를 변경한다. 배경 부분은 lama inpainting 기법을 통해서 얼굴이 없는 부분을 새롭게 배경 합성을 한다. 최종적으로 각각 만들어진 인물과 배경 사진들을 다시 합친 이미지는 기존의 hfgi editing을 통해 생성된 이미지보다 잔물결이 제거된 자연스러운 편집이 요구된다.

3.3. 시스템 설계

3.3.1 HFGI 네트워크 구조

HFGI 모델의 네트워크 구조는 그림 1에 묘사되어 있다. 원본 이미지가 fixed된 인코더에 들어오면 잠재 코드 W로 되고, fixed된 디코더를 통해 최초로 inversion된 이미지가 나온다. 이 이미지는 인코더-디코더로 압축 후 복원되는 과정에서 저차원의 잠재 벡터가 고주파의 섬세한 이미지 디테일을 놓치게 된다. 왜곡맵을 이용하여 놓친 정보를 획득한다. 왜곡맵과 원본 이미지 두개는 self-supervised learning하는 인코더-디코더 구조의 psp인코더 기반인 ADA 모듈에 들어가서 왜곡된 정렬을 다시 잡아준다. 이는 conv2d, batchnorm2d, PReLU로 이루어져 있다. 그 다음에는 Ec(Consulation encoder)에 들어간다. psp인코더 기반으로, 마찬가지로 conv2d와 batchnorm2d 그리고 PReLU로 이루어져 있다. HFGI를 training할 때 이 Ec가 훈련이 된다. Ec 에 들어오면 잠재 맵 C (18 *512 * H * W) 형태로 된다. 이때 잠재 코드 w (18*512)와 함께 consulation fusion이 fixed된 생성기의 첫 번째

레이어에 들어간다. 첫 번째 레이어에만 융합을 하는 이유는 과적합을 방지하기 위함이다. 이를 통해 생성된 이미지는 초반의 인코더-디코더를 통해 나온 것보다 정보를 덜 잃고 재현성과 편집성 모두 뛰어남을 보여준다. 그림2는 HFGI를 training할 때 이미지를 inversion하는 모습이다. 그림 3은 훈련된 Ec 인코더를 이용하여 추론을 하였을 때 다음과 같이 뛰어난 사진 재현을 보여준다. 그림 4에서는 HFGI에서 InterfaceGAN을 통해 잠재 벡터의 방향을 수정하여 이미지 속성 편집 중 'smile'을 변형했을 때의 결과이다. 그림 3과 그림 4를 통해 HFGI는 inversion과 smile 모두 뛰어난 이미지 편집을 보여준다. 하지만 그림 5를 보면 이미지 속성 편집 중 'pose'을 변형했을 때 이미지의 잔물결이 심함을 알 수 있다. 이때 HFGI에서 배포한 코드를 보면, pose를 변형하여 이미지 결과를 낼 수 있는 코드가 제외되어 있다. 따라서 코드를 수정 및 추가하여 pose 또한 변경하게 하였고, 극단적으로 시점을 변경하기 위해 edit degree는 3.0 (인물 기준 오른쪽으로 최대한 간 것) 으로 설정하였다.

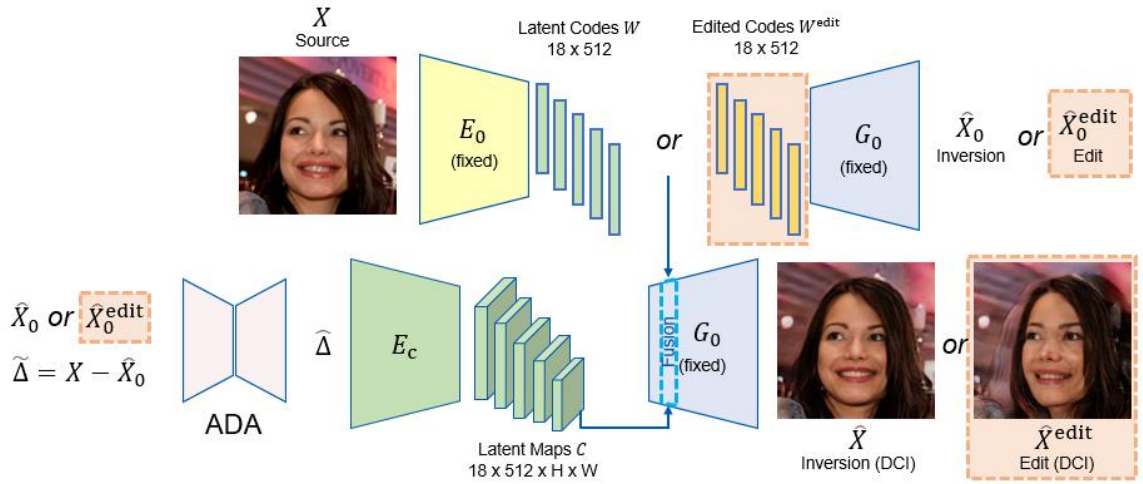

그림 1. HFGI 구조

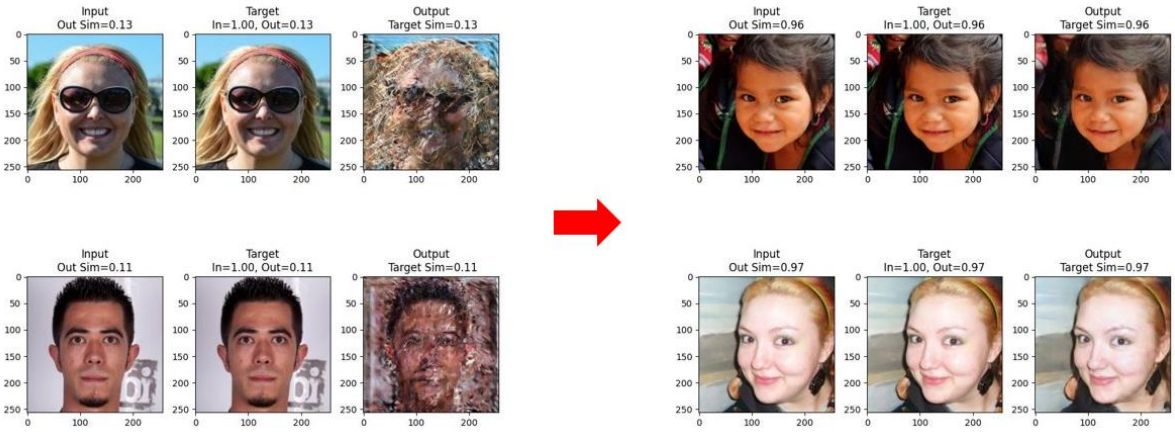

그림 2. HFGI training (inversion)

(좌) 원본 이미지 (우) inversion된 이미지

그림 3. HFGI inversion image

그림 4. HFGI 결과 - smile을 속성편집한 사진

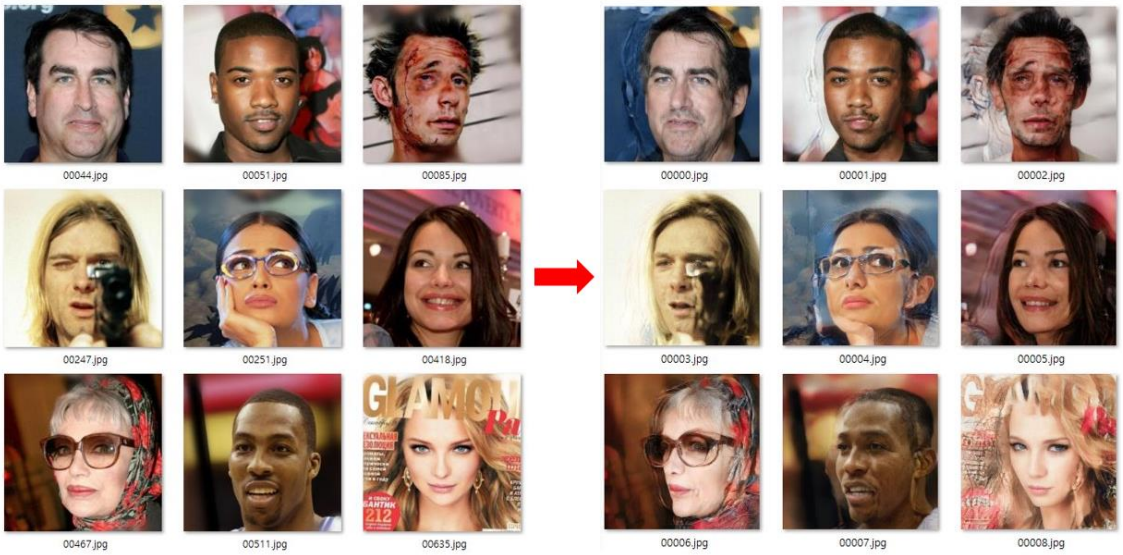

그림5. HFGI 결과 – pose를 속성편집한 사진

3.3.2 이미지 마스킹

그림 5와 같이 HFGI에서는 pose를 변경하였을 때 기존 인물의 사진 잔물결이 남아있는 문제점을 알 수 있다. 이를 제거하기 위해서 우선 원본 이미지와 편집된 이미지에서 각각 graphonomy을 이용하여 인물과 배경을 분리하는 작업을 추가하였다. hfgi/inference.py 아래에 graphonomy 코드 및 네트워크를 추가하여 이미지 마스킹을 하였다. 이때 모델은 graphonomy에서 제공한 pretrained 모델을 사용하였다. 그림 6, 7과 같이 이진 마스크 이미지를 생성하는 모습을 볼 수 있다. 여기에서 잔물결에 해당하는 부분은 두 이진 마스크를 비트연산 취했을 때 결과값으로 알아낼 수 있다. 잔물결 부분은 그림 8에 묘사되어 있다. 원본 이미지의 이진 마스크 사진을 이용하여 인물과 배경 영역을 구분하여 따로 저장한 결과는 그림 9와 그림10이다. 이때, 배경을 제외한 오직 인물 영역만을 hfgi의 pose 속성 편집을 하여 나온 결과는 그림11에 있다. 여기서 잔물결은 여전히 있지만 배경이 없기 때문에 정확히 어떤 식으로 잔물결이 남아있는지를 확인할 수 있다.

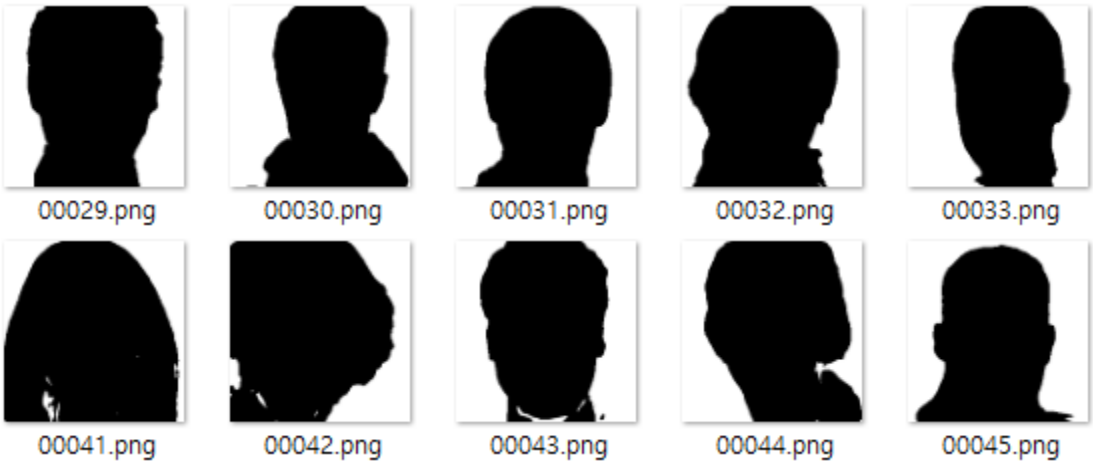

그림 6. 원본 이미지의 이진 마스크 사진

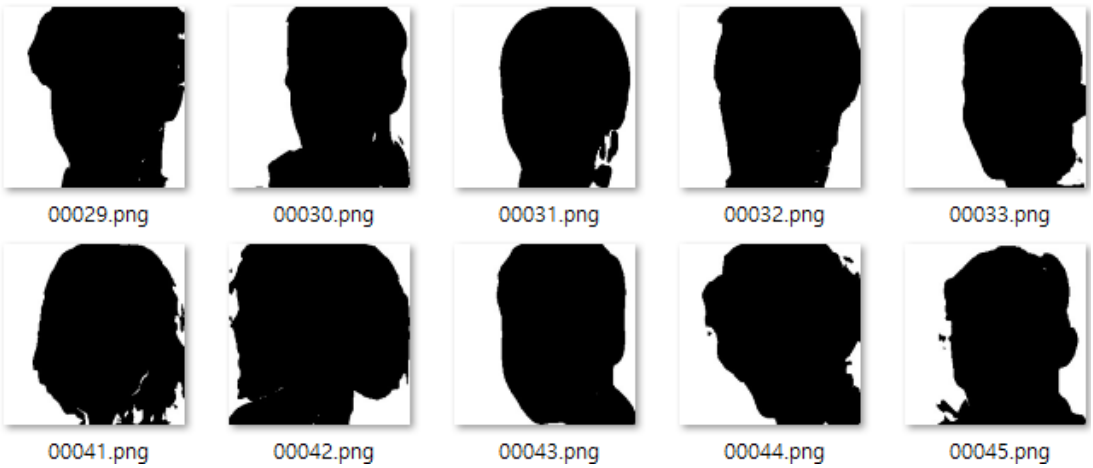

그림 7. 편집된 이미지의 이진 마스크 사진

그림 8. 잔물결에 해당하는 이진 마스크

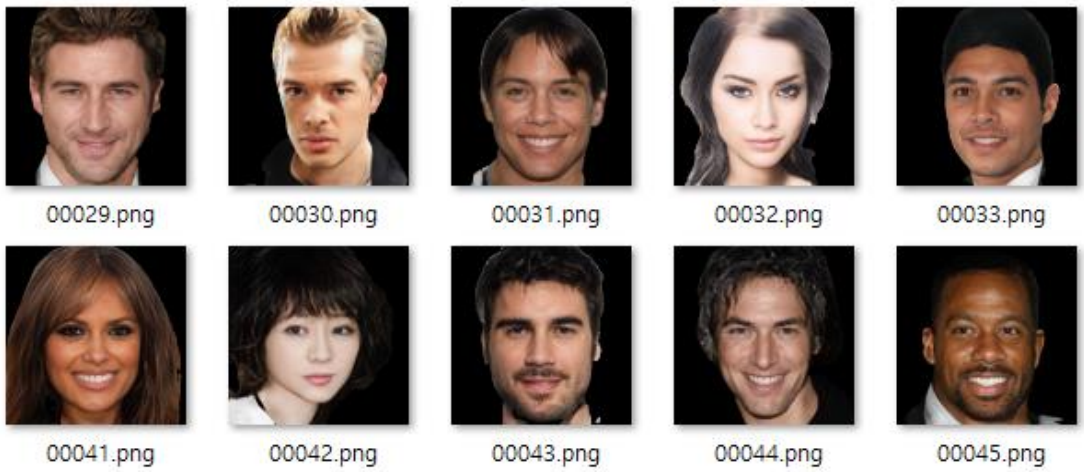

그림 9. 원본 이미지의 인물 영역

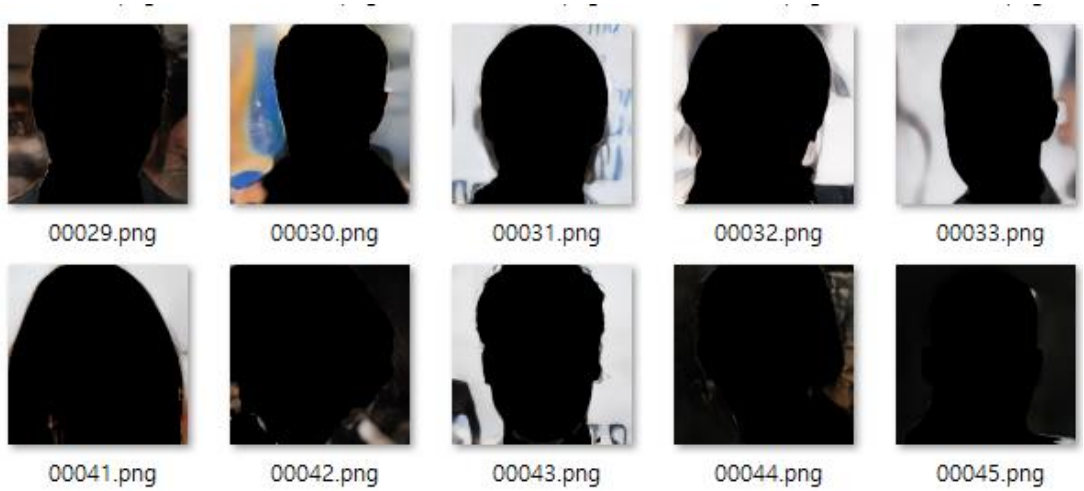

그림 10. 원본 이미지의 배경 영역

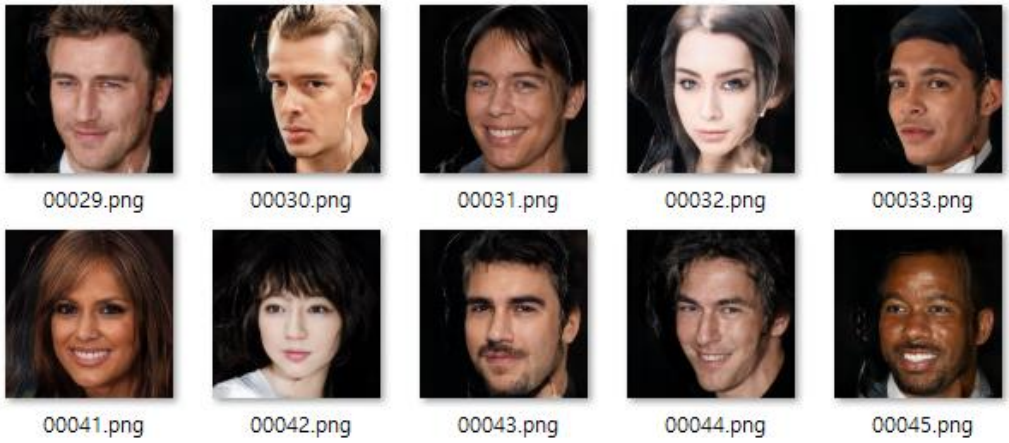

그림 11. 원본 이미지의 얼굴 영역을 HFGI pose 편집한 결과 사진

3.3.3 배경 합성 – inpainting 기법

그림 5와 같이 HFGI에서는 pose를 변경하였을 때 기존 인물의 사진 잔물결이 남아있는 문제점을 알 수 있다. 이를 해결하기 위해 image inpainting 기법을 이용하였다. Image inpainting은 이미지가 손상된 부분을 재구성하는 기술로 이미지 뿐만 아니라 비디오에서까지 확장되어 쓸 수 있는 딥러닝 기술이다. 인페인팅은 손상/누락된 픽셀들을 채우는 것이 목적으로, 현재까지 많은 기법들이 발전되고 있다. 그중에서 올해 2022년에 나온 LaMA Image Inpainting을 이용하여 문제되는 잔물결 부분의 배경을 inpainting 해보았다. 결과는 4.3 에 명시되어있다.

4. 프로젝트 결과

4.1 개발 환경

프로젝트 진행은 학교에 신설된 서버gpu를 이용하였다. sftp 접속을 통해 접속하여 로컬환경에 비해 더 좋은 환경에서 딥러닝을 할 수 있었다. 편집기는 vs code로, 특정 포트를 통해 교외접속해서 실시간으로 코드를 수정 및 업로드를 하였다.

- OS : Ubuntu 20.04.5 LTS (GNU/Linux 5.4.0-131-generic x86_64)
- VGA : NVIDIA GeForce RTX 3090
- CPU : Intel(R) Xeon(R) E-2334 CPU @ 3.40GHz
- VS Code (Terminal : Git Bash)

4.2 연구 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].

4.3 연구 결과

왼쪽은 HFGI에서 pose를 editing한 사진이고, 오른쪽은 masking 및 inpainting기법을 적용한 결과이다. 왼쪽 사진과 잔물결 영역의 이진 마스크 사진을 lama에 넣었을 경우는 다음과 같다.

5. 결론

5.1. 기대효과

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 잔물결이 남는 치명적인 문제점을 발견하였다. 이를 해결하기 위해 graphonomy를 이용하여 이미지의 배경과 인물을 따로 이진마스크 생성하여 취하였다. 인물은 기존대로 hfgi의 InterfaceGAN에 따라 latent code의 방향을 수정하여 속성 편집을 하였다. 마지막으로 배경과 인물을 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 개선을 보여준다.

5.2. 추후 연구 방향

이번 프로젝트에서는 잔물결의 영역을 구하는 방식으로 graphonomy를 이용하여 이진 마스크 이미지들을 생성하여 비트 연산을 통해 얻는 방식을 취하였다. 잔물결의 영역에 해당하는 이진 마스크 사진을 이용하여 기존에는 없었던 배경 부분에서 inpainting을 통해 새롭게 합성하는 방식을 취하였다. 하지만 이 잔물결의 영역을 스스로 masking할 수 있는 모델을 만들어 추가한다거나, 잔물결이 있는 부분에 대하여 HFGI에서 fusion이 이루어지는 구간에서 배경영역이 그쪽까지 보간되어 합성할 수 있는 새로운 방법이 있을 것이다. 또는 ADA모듈을 좀 더 강화하여 추가적인 학습을 통해 시점 변경이 큰 속성 편집 이미지에서도 큰 문제가 없이 돌아갈 수 있게 하는 방법이 있을 것이다. 추후에는 잔물결에 대한 부분을 더 자연스럽게 하기 위하여 추가적인 모델 설계를 하여 HFGI를 보다 더 개선할 수 있을 것으로 예상된다.

6. 참고문헌

[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.

[11] Gong, Ke, et al. "Graphonomy: Universal human parsing via graph transfer learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[12] Suvorov, Roman, et al. "Resolution-robust large mask inpainting with fourier convolutions." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.