확산 기반 얼굴 교환 모델의 얼굴 세부 특징 보존을 통한 성능 개선

최재웅, 이재구*
국민대학교 일반대학원 컴퓨터공학과

* jaekoo@kookmin.ac.kr

Improving the Performance of a Diffusion-Based Face Swapping Model Through Preserving Fine Facial Features

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

요 약

얼굴 교체(face swap) 과업은 원본 얼굴 사진의 특징을 추출하여 목표 얼굴 사진과 교체하며, 원본 얼굴 사진의 특징을 최대한 보존하는 것을 목표로 한다. 기존에는 GAN 기반 얼굴 교체 과업이 많이 이루어져 왔으나, 최소극대화(minimax) 최적화의 불안정성 문제로 인해 확산 모델(diffusion model)이 새로운 생성자로서 제안되어왔다. 새로운 생성자로서 확산 모델 또한 얼굴 교환 과업에도 적용되었으나, 교체 시 눈, 코, 입과 같은 전체적인 특성은 잘 보존하되 쌍꺼풀, 입술과 같은 세부적 얼굴 특성은 잘 보존하지 못하는 것을 실험적으로 확인하였다. 그렇기에 본 논문은 얼굴의 세부적인 특성을 추출하는 추가적인 인코더(encoder)를 사용하였으며, 두가지 정량적 평가를 통해 성능이 개선됨을 입증하였다.

Ⅰ. 서 론

최근, 생성형 인공지능(generative AI)은 컨텐츠로써 많이 사용되어 왔다. 그 중 사진 편집이나 영상 편집은 사용자들의 큰 관심사 중 하나이다.

본 논문에서 제안하는 모델은 생성형 인공지능 중, 얼굴 교체(face swap) 과업을 수행한다. 얼굴 교체 과업은 그림 1 과 같이 목표 얼굴 사진(target face image)의 특징을 원본 얼굴 사진(source face image)의 특징으로 교체하며 원본 얼굴 사진의 특징을 최대한 보존하는 것을 목표로 한다.

기존의 얼굴 교체 연구는 GAN 기반으로 하는 연구[1]가 주를 이루었다. 그러나 GAN 기반의 연구는 많은 연구에도 불구하고, 생성자(generator)와 판별자(discriminator)의 최소극대화(minimax)의 최적화 문제로 인해 불안정하다는 것을 여러 실험을 통해 확인되어왔다[2]. 그렇기에 본 논문은 확산모델(diffusion model)을 기반으로 얼굴 교체 과업을 수행한다. 확산 모델을 기반으로 하는 기존 연구[2]는 샘플링(sampling)하는 과정에서 사전 학습된(pretrained) 인코더(encoder)를 지도(guidance)로 사용하여 잡음 제

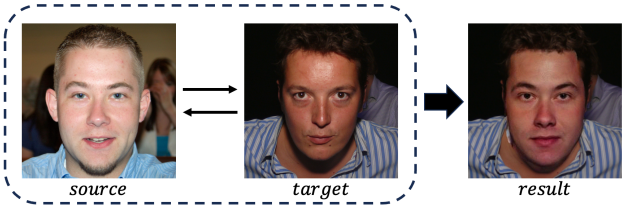

그림 1. 얼굴 교환 과업

거(denoising) 과정을 제어한다. 지도는 특정 특징을 결과에 추가하기 위해 사용되며, 특히 원본 얼굴의 특징을 추가하기 위해 먼저 ArcFace[3]로 원본 얼굴의 특징을 추출한다. ArcFace 는 얼굴의 특징을 추출한 뒤 벡터(vector)로 변환하는 과업을 수행하며, 이를 통해 얼굴 인식(face recognition)과 얼굴 복원(face reconstruction)을 수행한다. 그러나 이전의 GAN 기반 얼굴 복원(face reconstruction) 연구[4]를 통해 ArcFace 만으로는 세부적인 특성인 고주파 특성(high-frequency attribute)을 복원하기 어렵다는 것을 입증했다. 그렇기에 본 논문은 잡음제거를 제어하는 과정에서 원본 얼굴 사진의 고주파 특성 보존을 위해 지도를 추가하여 성능 개선방법을 제안한다.

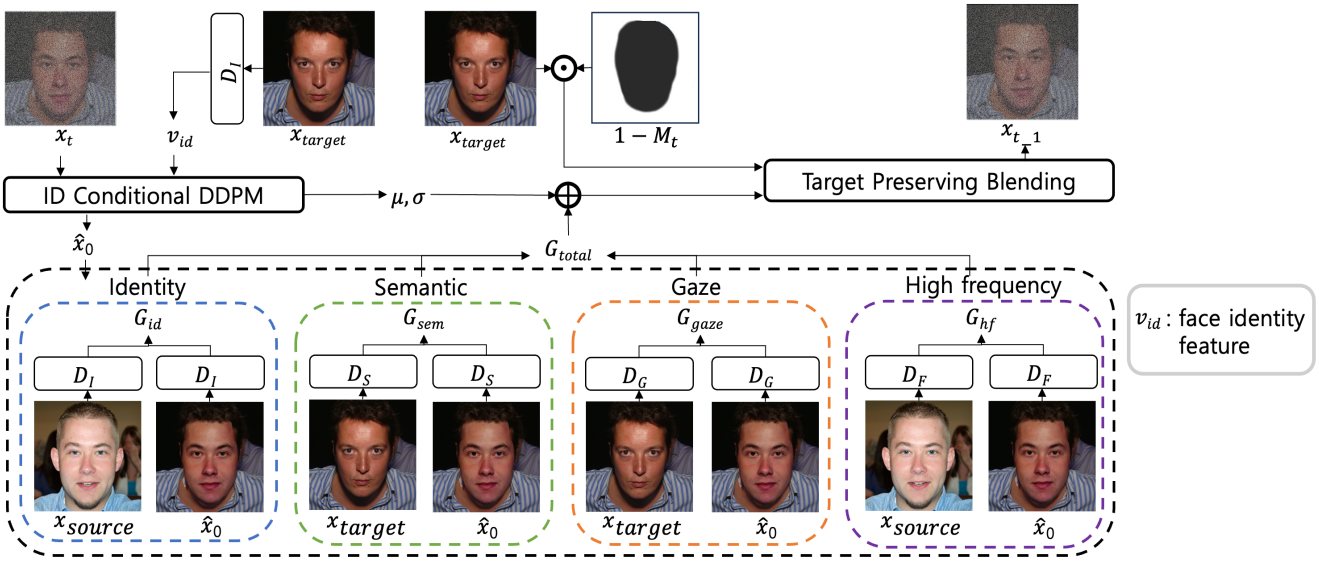

그림 2. 제안 모델의 개요

원본얼굴사진 목표얼굴사진 기존모델결과 제안모델결과

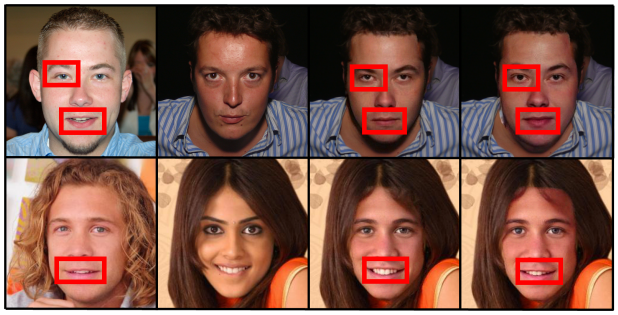

그림 3. 두 모델의 결과 비교

Ⅱ. 본 론

본 논문은 확산을 기반으로 하는 얼굴 교체 모델의 개선을 제안한다. 먼저 본 논문은 사전 학습된 ID Conditional DDPM[2]을 통해 무작위 잡음으로부터 목표 얼굴 사진 x_{target}의 얼굴 특징을 보존하도록 샘플링한다.

다음, ID Conditional DDPM 의 샘플링 과정(sampling process)에서 지도(guidance)를 통해 결과를 제어한다. 단, 이때 손실 계산을 위해 $\hat{x}_0$ 를 사용하는데, $\hat{x}_0$ 는 t 시점에서 잡음이 낀 사진에서 잡음을 완전히 제거한 결과이다. 이 과정을 통해 샘플링의 결과를 변화된다.

먼저, 첫번째 지도인 정체성(identity) 지도 G_{id}를 통해 $\hat{x}_0$ 의 정체성이 원본 얼굴사진 x_{source} 의 정체성과 유사해지도록 샘플링한다. 정체성을 추출하기 위해 정체성 추출기[4] D_I를 사용한다.

$$G_{id} = 1 - \cos\left(D_I(x_{source}), D_I(\hat{x}_0)\right) \quad (1)$$

다음, 두번째 지도 $G_{semantic}$ 을 통해 $\hat{x}_0$ 가 x_{target} 의 의미론적(semantic) 정보를 담도록 제어한다. 이를 위해 얼굴 분할기[6] D_S로 $\hat{x}_0$ 의 얼굴 분할 결과와 x_{target}의 얼굴 분할 결과가 유사해지도록 샘플링한다.

$$G_{semantic} = \left\|D_S(x_{target}) - D_S(\hat{x}_0)\right\|_2^2 \quad (2)$$

다음, 세번째 지도 G_{gaze} 를 통해 $\hat{x}_0$ 가 x_{target} 의 시선(gaze)과 동일하도록 제어한다. 이를 위해 $\hat{x}_0$ 와 x_{target} 가 시선 추출기 D_G 를 거친 뒤 둘의 임베딩(embedding) 결과가 유사해지도록 샘플링한다.

$$G_{gaze} = \left\|D_G(x_{target}) - D_G(\hat{x}_0)\right\|_2^2 \quad (3)$$

그러나 x_{source} 의 세부적인 특징인 고주파 특성을 담기에는 구조적 유사성에 집중하는 G_{id} 만으로는 부족하다[4]. 따라서 본 논문은 세부적인 특징을 잘 샘플링 하는 고주파(high frequency) 지도 G_{hf} 를 제안한다. D_F 는 세부적인 특징을 추출하는 EleGAN-t 인코더[7] 이며, x_{source} 의 세부적인 특성과 x_{target} 의 세부적인 특성이 유사해지도록 샘플링한다. .

$$G_{hf} = \left\|D_F(x_{source}) - D_F(\hat{x}_0)\right\|_1^1 \quad (4)$$

최종적으로 모든 지도를 더한다.

$$G_{total} = G_{id} + G_{semantic} + G_{gaze} + G_{hf} \quad (5)$$

또한 확산 모델의 경우, 잡음제거 과정에서 배경의 보존을 실패한다. 따라서 x_{target}에서 얼굴 마스크(mask) M_t 를 사용하여 배경만 남긴 후 Target Preserving Blending 에 입력으로 넣어 배경을 보존한다.

$$x_{t-1} = \hat{x}_{t-1} \odot M_t + x_{t-1,target} \odot (1 - M_t) \quad (6)$$

이를 통해 확산 모델의 추출과정에서 지도를 통해 손실에 따른 제어된 사진을 생성한다.

Ⅲ. 실험

제안된 모델의 ID Conditional DDPM 은 FFHQ[8]를 통해 학습된 확산 모델이며, 평가 시 Celeba-HQ[9]를 원본 얼굴 사진과 목표 얼굴 사진으로서 사용하였다. 그림 3 의 결과를 통해 기존 모델 결과가 제안 모델 결과에 비해 쌍꺼풀, 입술, 이빨과 같은 세부적인 부분을 더 잘 표현함을 정성적으로 확인하였다. 또한 세부적인 특성인 고주파 특성을 잘 표현하였는지 정량적으로 확인하기 위해 얼굴 유사도 측정 R(face similarity) [2]과 SSIM[4] 점수를 사용하였다. R 은 $source$ 와 $result$ 가 유사하며, $target$ 과 유사하지 않을수록 수치가 높다. 얼굴 유사도 측정은 식 7 을 통해 계산하였다.

표 1. 두 모델의 정량적 지표

	R(face similarity) (↑)	SSIM(↑)
기존 모델[3]	48.9	0.316
제안 모델	50.5	0.325

$$R := \frac{D(result,\ source)}{D(result,\ source) + D(result,\ target)} \quad (7)$$

D 는 정체성 추출기를 거친 두 사진의 임베딩 간 코사인 거리(cosine distance)이다. 표 1 을 통해 제안된 모델의 유사도 측정 수치가 1.6 높은 것이 확인됨으로써 제안모델이 원본 얼굴 사진의 얼굴 특징을 더 잘 보존한다는 것을 입증하였다. 또한 SSIM 점수는 두 사진 간 세부적 특징인 고주파 특징을 얼마나 더 잘 보존되어 있는지 확인하는 지표이다. 제안 모델이 0.09 높게 나온 결과를 통해 세부적 특징을 더 잘 보존함을 입증하였다.

Ⅳ. 결 론

기존 확산 기반 얼굴 교환 과업은 샘플링 과정에서 지도를 통해 잡음 제거 과정을 제어한다. 그러나 기존 모델의 경우, 세부적인 특성을 잘 표현하지 못하였다. 이를 개선하기 위해 본 논문의 제안 모델은 세부적인 특징을 추출하는 EleGAN-t 인코더를 지도로서 사용하였으며, 정성적, 정략적 평가를 통해 제안 모델의 성능이 더 나음을 입증했다.

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] Chen, Renwang, et al. "Simswap: An efficient framework for high fidelity face swapping." *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.

[2] Kim, Kihong, et al. "Diffface: Diffusion-based face swapping with facial guidance." *arXiv preprint arXiv:2212.13344* (2022).

[3] Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[4] Moon, Seung-Jun, Chaewon Kim, and Gyeong-Moon Park. WaGI: Wavelet-based GAN Inversion for Preserving High-frequency Image Details." *arXiv preprint arXiv:2210.09655* (2022).

[5] Park, Seonwook, et al. "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings." Proceedings of the 2018 ACM symposium on eye tracking research & applications. 2018.

[6] Yu, Changqian, et al. "Bisenet: Bilateral segmentation network for real-time semantic segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.

[7] Yang, Chenyu, et al. "Elegant: Exquisite and locally editable gan for makeup transfer." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[8] 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.

[9]FaceSwap,https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski, accessed: 2022-02-14.