자기 지도 학습을 통한 고해상도 얼굴 영상 복원

조병호, 박인규
인하대학교 정보통신공학과
byunghojo12@gmail.com, pik@inha.ac.kr

Face Super Resolution using Self-Supervised Learning

Byung-Ho Jo, In Kyu Park
Department of Information and Communication Engineering, Inha University

요 약

본 논문에서는 GAN 과 자기 지도 학습(self-supervised learning)을 통해 입력 얼굴 영상의 공간 해상도를 4 배 증가시키는 기법을 제안한다. 제안하는 기법은 변형된 StarGAN v2[1] 구조의 생성자와 구분자를 사용하여 저해상도의 입력 영상만을 가지고 학습 과정을 거쳐 고해상도 영상을 복원하도록 자기 지도 학습을 수행한다. 제안하는 기법은 복원된 영상과 고해상도 영상 간의 손실을 줄이는 지도 학습이 가지고 있는 단점을 극복하고 입력 영상만을 가지고 영상 내부에 존재하는 특징을 학습하여 얼굴 영상에 대한 고해상도 영상을 복원한다. 제안하는 기법과 Bicubic 보간법과의 비교를 통해 우수성을 검증한다.

1. 서론

저해상도 얼굴 영상에서 고해상도 얼굴 영상 복원은 초해상화(Superresolution) 연구에 있어 얼굴 영역에 대한 특화 분야로, 2D 얼굴 영상 분석 및 3 차원 얼굴 복원을 위한 중요한 기법이다. 최근 CNN 의 발전으로 저해상도 영상과 고해상도 영상의 쌍을 이용한 지도 학습(supervised learning)을 통해 우수한 결과를 보이는 다양한 연구가 진행되었다. 기존의 연구는 주로 매우 낮은 공간 해상도(32x32)의 얼굴 영상을 256x256 의 해상도로 복원하도록 진행되었다. 하지만 정확한 얼굴 영상 분석 및 3 차원 얼굴 복원을 위해서는 더 높은 해상도의 얼굴 영상이 요구된다. 그러나 지도 학습을 통해 더 높은 고해상도 영상 복원을 위해서는 기존 연구에 쓰이는 데이터보다 더 높은 해상도의 데이터셋 구축이 필요하다. 이는 학습 과정에서 고비용의 데이터셋과 하드웨어 구축을 요구하는 점과 더불어, 높은 복잡도(curse of dimensionality)를 가지는 단점이 존재한다. 본 논문에서는 문제점을 해결하고자 변형된 StarGAN v2[1]구조와 ZSSR[2]에서 제안한 깊은 내부 학습(deep internal learning) 및 자기 지도 학습을 바탕으로 256x256 해상도의 입력 영상만을 가지고 학습 과정을 거쳐 1024x1024 의 고해상도 얼굴 영상을 복원한다.

2. 제안하는 기법

제안하는 기법의 구조는 그림 1 에 도식하였다. 본 논문은 자기 지도 학습 기법을 수행하기 위해, 훈련 시 입력 영상을 GT 로 취급하고, 입력 영상의 공간 해상도를 4 배 줄여 이를 훈련 시 입력 영상으로 만들었다. 본 논문은 이와 같이 기존 연구에서 진행된 지도 학습처럼 저해상도와 고해상도 영상의 쌍(64→256)을 만들어 학습을 진행하였다. 추가적으로, 입력 영상에 대해 다양한 스케일(16→64, 32→128)의 쌍을 만들어 모델이 얼굴 영역의 추상적인 정보와 디테일한 정보를 학습하도록 수행하였다. 또한, 90, 180, 270 회전, 가우시안 노이즈, 가우시안 블러, JPEG 압축 손실 노이즈 중 하나를 임의로 입력 영상에 적용하였다. 이를 통해 본 논문에서 제안하는 모델은 얼굴 영상 내부에 존재하는 얼굴 영역의 특징을 학습하며 얼굴 영역에 대한 복원을 학습할 수 있도록 하였다. 본 논문에서 제안한 자기 지도 학습을 통해 제안하는 모델은 지도 학습 보다 더 사실적인

얼굴 영상을 복원한다.

2.1 제안하는 모델

제안하는 모델은 입력 영상의 공간 해상도를 4 배로 키우는 생성자 네트워크와 입력 영상의 진위를 판별하는 구분자 네트워크를 포함한다. 제안하는 모델은 안정된 학습을 위해 adversarial loss 를 LSGAN[4]의 평균 제곱 오차 함수(mean squared error function)로 채택하였다.

제안하는 모델의 생성자는 임의로 적용된 노이즈와 다양한 공간 해상도의 영상를 입력으로 받아 원본 영상 내부에 존재하는 통계 정보를 학습하며 고해상도 얼굴 영상을 복원하도록 한다. 구분자는 다양한 스케일의 영상을 입력 받으며 GT 영상과 복원된 영상의 진위를 패치(patch) 단위로 판별하도록 학습을 하였다.

2.2 손실 함수

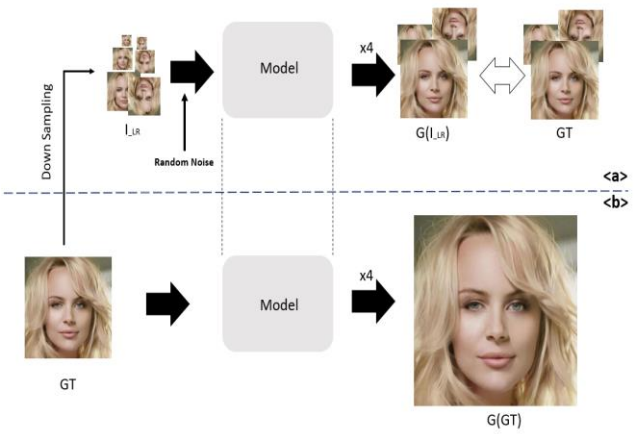

그림 1. 제안하는 기법의 구조. a: 훈련 과정, b: 테스트 과정

제안하는 모델에서 사용한 생성자 G 와 구분자 D 에 대한 목적 함수 L_G와 L_D 는 다음과 같다.

$$minL_G = E_{I_{HR},GT}\left[\, \|G(I_{LR}) - GT\|_{L_1} \,\right] + \frac{1}{2}E_{I_{HR},GT}\left[\, \left(D(G(I_{LR})) - 1\right)^2 \,\right] + E_{I_{HR},GT}\left[\, \Sigma_i \| \, f_i(G(I_{LR})) - f_i(GT) \, \|_{L_1} \,\right]$$

$$minL_D = \frac{1}{2}E_{GT,Real}[(D(GT) - 1)^2] + \frac{1}{2}E_{I_{HR},Fake}\left[D(G(I_{LR}))^2\right]$$

L_G 는 입력 영상 I_{LR} 로부터 복원된 얼굴 영상 I_{HR} 과 GT 영상과의 오차를 픽셀 단위의 L_1 거리, 구분자에 대한 adversarial loss 와 판단 손실(perceptual loss)를 나타낸다. f는 ImageNet 데이터셋에 학습된 VGG-19 모델의 i 번째의 층에서 추출된 특징을 나타낸다.

L_D 는 다양한 공간 해상도의 영상을 입력으로 받아 원본 영상의 여부를 판별하는 adversarial loss 를 나타낸다.

3. 실험 결과

본 논문에서는 실험을 위해 CelebA-HQ 데이터셋[3]을 사용하였다. 총 30,000 장 중의 얼굴 영상 중에 3,998 장을 학습에 사용하였으며 영상에 4 배의 다운 샘플링을 적용하여 훈련과 테스트를 진행하였다. 또한 본 논문은 자기 지도 학습 기법을 제안하므로 학습에 사용된 데이터셋이 테스트 데이터로 사용되지만, 저해상도 얼굴 영상 복원의 일반화 역량을 실험하기 위하여 훈련에 사용하지 않은 5,000 장을 추가로 테스트하였다. 제안하는 모델은 8의 배치 크기를 갖도록 하여 GTX 2080 TI GPU 에서 훈련되었다. 정량적 평가를 위해 PSNR, SSIM 과 FID 를 사용하여 Bicubic 보간법을 영상에 적용한 결과와 비교하였다. 표 1 과 2 를 통해 제안하는 기법이 PSNR 과 SSIM 결과는 Bicubic 에 비해 낮은 것을 알 수 있다. 하지만 얼굴 영상 복원 결과의 평가에 있어 픽셀 단위 기반의 비교보다는 원본 영상과 얼마나 perceptual 하게 같은지 평가하는 FID 가 더 적절하다. 제안하는 기법이 Bicubic 에 비해 월등히 우수한 FID 값을 나타내며 더 사실적인 얼굴 영상을 복원함을 알 수 있다. 정성적 평가 결과는 그림 2, 3 에 보였으며 제안하는 기법이 얼굴 영역을 더 사실적으로 복원함을 보인다. 그림 4 를 통해 지도 학습 기법으로 (256→1024) 고해상도의 얼굴 영상을 복원하는 연구(PanGAN)[5]와 비교하였다. 본 논문이 제안하는 자기 지도 학습 기법이 지도 학습 기법에 비해 원본 영상의 얼굴 영역의 특징을 학습하며 GT 와 더 가까운 고해상도의 영상으로 복원함을 알 수 있다.

	Bicubic	Ours
PSNR	32.51	31.65
SSIM	0.8535	0.8347
FID	19.17	4.31

표 1. 훈련 데이터 정량적 평가 결과

	Bicubic	Ours
PSNR	36.96	33.32
SSIM	0.9514	0.8974
FID	9.30	3.29

표 2. 테스트 데이터 정량적 평가 결과

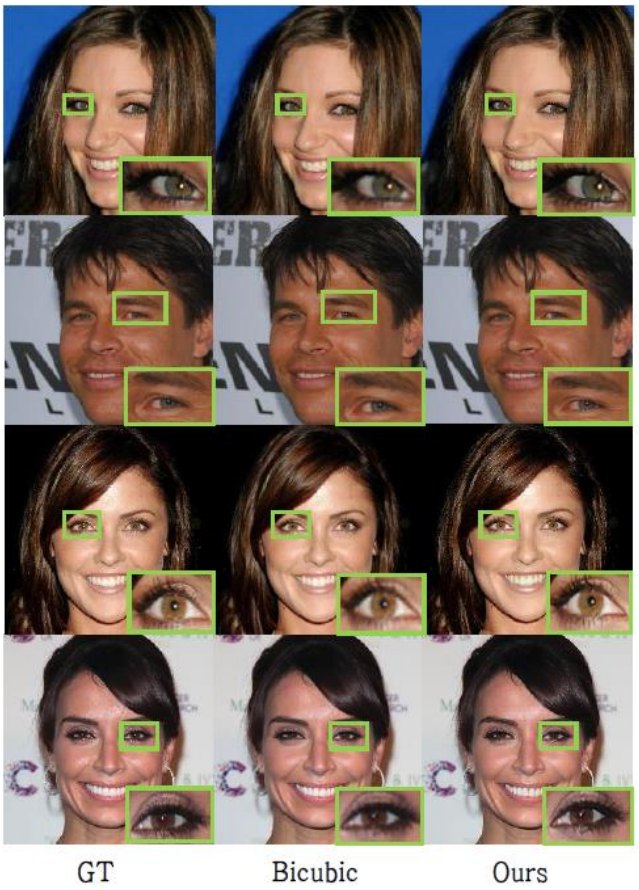

그림 2. 정성적 평가 결과. 영상 해상도는 1024x1024 이다.

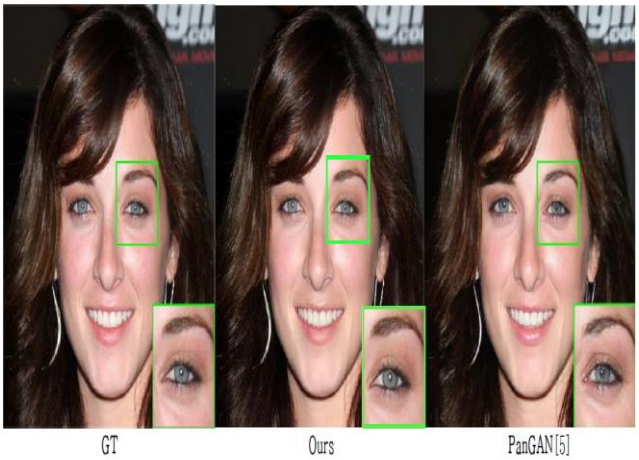

그림 3. 정성적 평가 결과.

4. 결론

본 논문에서는 자기 지도 학습을 통해 입력 영상만을 사용하여 고해상도의 얼굴 영상을 복원하는 기법을 제안하였다. 실험 결과를 통해 기존의 영상 보간법보다 정량적으로, 기존의 지도 학습 기법보다 정성적으로 높은 성능을 나타내는 것을 알 수 있다. 본 논문에서 제안하는 기법은 기존 지도 학습의 단점을 극복하고 한 개의 학습 영상에 대해서만 복원(an image specific model)을 하는 ZSSR[2] 과 달리, 학습에 쓰이지 않은 얼굴 영상에 대해서도 사실적인 얼굴 영상을 복원하였다.

5. 감사의 글

이 논문은 2020 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019R1A2C1006706).

6. 참고문헌

[1] Y. Choi, Y. Uh, J. Yoo and J. Ha, “StarGAN v2: Diverse Image Synthesis for Multiple Domains,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2020

[2] A. Shocher, N. Cohen and M. Irani, ““Zero-Shot” Super-Resolution using Deep Internal Learning,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018

[3] T. Karras, T. Aila, S. Laine and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” International Conference on Learning Representations, 2018

[4] X. Mao, Q. Li, H. Xie and R. Y.K., et al., “Least Squares Generative Adversarial Networks,” Proc. of IEEE International Conference of Computer Vision, 2017

[5] L. Wong, D. Zhao, S. Wan and B. Zhang, “Perceptual Image Super-Resolution with Progressive Adversarial Network,” arXiv:2003.03756v4[Preprint], 19 Mar 2020. Available from: arxiv.org/abs/2003.03756