회귀 기반 신경망과 적대적 생성망을 함께 활용한 영상 복원

이화윤 [O,1], 강경국 [2], 이형민 [1], 백승환 [1,2], 조성현 [1,2]
포항공과대학교 {인공지능대학원 [1], 컴퓨터공학과 [2]}
{hwayoon2, kkang831, hmin970922, shwbaek, s.cho}@postech.ac.kr

요 약

최근 딥러닝을 이용하여 영상을 복원하는 다양한 방법들이 제안되었다. 이들은 크게 회귀 기반 방법과 생성 기반 방법으로 나눌 수 있다. 픽셀단위의 손실을 최소화하여 정답 영상의 구조 복원을 목표로 하는 회귀 기반 방법은 높은 품질의 사실적인 영상 복원에 어려움을 겪는다. 적대적 생성망을 선행 지식으로 이용하는 생성 기반 방법은 높은 품질의 영상을 복원하지만 원본 이미지의 구조를 정확하게 복원하지 못한다. 본 논문에서는 기존의 회귀 기반 신경망을 이용해 정확한 구조를 복원하면서 적대적 생성망을 함께 활용하여 보다 사실적인 영상을 복원하는 방법을 제안한다.

1. 서론

영상 복원은 블러, 노이즈, 낮은 화질 등에 의해 열화 된 영상으로부터 열화가 제거된 영상을 복원하는 것을 목표로 한다. 최근 딥러닝이 발전하면서 영상 복원을 위한 다양한 인공 신경망들이 [1,2] 제안되었다.

오랜 시간을 거쳐 다양한 신경망 구조들이 연구되어온 회귀 (Regression) 기반 방법은 주로 합성 곱 신경망을 이용하여 열화 된 영상으로부터 정답 영상의 구조를 복원하는 것을 주 목표로 한다. 이 방법은 대부분 픽셀 간의 평균 제곱 오차 (Mean Squared Error) 또는 평균 절대 오차 (Mean Absolute Error)를 최소화하도록 학습되어 정확한 구조를 복원할 수 있다. 하지만 입력 영상에 상응할 수 있는 다양한 결과들의 평균을 생성하게 만드는 손실 함수의 특성 때문에 낮은 품질의 흐릿한 영상을 생성한다.

이를 해결하고자 등장한 방법 중에 하나인 생성 (Generative) 기반 방법은 적대적 생성망 (Generative Adversarial Network)의 정보를 선행 지식으로 이용하는 방법이다. 적대적 생성망은 영상을 생성하는 생성자 (Generator)와 생성된 영상과 실제 영상을 구분하는 판별자 (Discriminator)로 구성된다. 생성자는 잠재 벡터로부터 영상을 생성하는데 판별자를 속일 수 있도록 학습되어 실제 영상들과 비슷한 높은 품질의 영상을 생성한다. 판별자는 입력 받은 영상이 실제 영상인지 생성자에 의해 만들어진 영상인지를 판별하도록 학습된다. 두 신경망의 적대적인 학습을 통해 생성자는 판별자를 속일 수 있도록 더 사실적인 영상을 생성하게 된다.

생성 기반 방법은 이러한 적대적 생성망의 사실적 영상 생성 능력을 활용하여 복원 영상을 합성한다. 이 방법은 주로 주어진 열화 영상을 인코더를 사용하여 적대적 생성망의 잠재 공간으로 사상하는 인버전을 통해 생성자가 정답 영상과 유사한 영상을 생성해 낼 수 있는 잠재 벡터를 찾는다. 이후 생성자가 만들어낸 영상을 보완하기 위해 입력된 열화 영상으로부터 형태 정보를 추출하여 이를 조정하는 과정을 거친다.

생성 기반 방법은 생성자가 만들어낸 영상이 기반이 되기 때문에 실제와 비슷한 높은 품질의 영상 복원이 가능하다. 하지만 정답 영상의 형태를 정확히 복원하지 못하는 경우가 존재한다. 이는 적대적 생성망의 잠재 공간이 한정적이기 때문에 정답 영상이 생성자가 표현할 수 있는 범위를 벗어날 수 있기 때문이다. 이를 보완하기 위해 영상을 조정하는 추가적인 과정을 거치지만 열화 영상으로부터 형태 정보를 얻기 때문에 정확한 형태 정보를 얻기 어렵다. 생성 기반 방법은 또한 신경망 구조가 고정되어 오랜 시간 연구된 다양한 회귀 기반 신경망 구조를 사용할 수 없다는 한계점을 가진다. 예를 들어 강한 모션 블러와 같이 복원하기 어려운 열화를 복원하기 위해 잘 고안된 회귀 기반 신경망들이 존재하지만 생성 기반 방법들은 이를 활용할 수 없다.

본 논문에서는 정교하게 고안된 회귀 기반 생성망을 활용하여 다양한 열화에 대해 정확한 복원이 가능하면서 적대적 생성망을 선행 지식으로 활용하여 사실적이고 높은 품질의 영상을 복원하는 방법을 제안한다. 본 논문에서 제안하는 방법은 구체적으로 열화 된 이미지를 회귀 기반 신경망을 거쳐 복원하여 정확한 구조의 영상을 얻는다. 이를 적대적 생성망의 잠재 공간으로 사상하여 좋은 품질의 인버전 영상을 얻는다. 최종적으로 문맥 손실 함수를 이용해 학습된 퓨전 신경망을 이용하여 정확한 구조를 가진 낮은 품질의 복원 영상과 좋은 품질의 부정확한 인버전 영상으로부터 정확하고 좋은 품질의 영상을 얻는다.

본 논문에서 제안한 구조를 회귀 신경망 NAFNet [1]을 사용해 모션 블러에 적용하는 실험을 진행했다. 이를 통해 본 논문에서 제안한 방법이 기존의 회귀 기반 방법보다 좋은 품질의 영상을 생성하고 생성 기반 방법보다 정확한 복원이 가능함을 정량적, 정성적으로 보인다.

2. 회귀 기반 신경망과 적대적 생성망을 함께 활용한 영상 복원

2.1 에서는 회귀 기반 신경망과 적대적 생성망을 함께 활용하는 전체 구조에 대해 설명한다. 2.2 에서는 적대적 생성망을 선행지식으로 활용하기 위해 고안한 인버전 인코더를 설명한다. 2.3 에서는 회귀 기반 신경망과 적대적 생성망의 정보들을 조합하는 퓨전 신경망에 대해 설명한다. 2.4 에서는 각 신경망을 학습시키는 방법에 대해 설명한다.

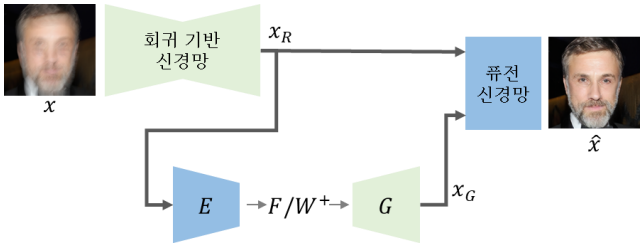

그림 1. 회귀 기반 신경망과 적대적 생성망을 함께 활용하는 영상 복원 프레임 워크 전체 구조

2.1 전체 구조

그림 1 은 본 논문에서 제안하는 방법의 전체 구조이다. 먼저 열화 된 영상 (x)를 입력으로 받아 회귀 기반 신경망을 거쳐 복원된 영상 (x_R)을 얻는다. 정답 영상의 구조를 정확하게 복원하도록 학습된 회귀 기반 신경망을 사용하여 정확한 구조를 복원하지만 회귀 기반 신경망의 단점인 선명하지 못한 낮은 품질의 영상이 생성된다. (그림 2 의 (a))

다음으로 회귀 기반 신경망 복원 영상 (x_R)을 적대적 생성망의 잠재 공간으로 사상하여 인버전 된 영상(x_G)를 얻는다. 적대적 생성망으로는 뛰어난 영상 합성 능력을 가진 StyleGAN2 [6]를 사용했다. 인버전 된 영상은 잠재 공간의 한정된 표현력 때문에 정답 영상의 구조를 정확하게 복원해내지는 못한다. 하지만 판별자를 속일 수 있도록 학습된 생성자를 선행 지식으로 이용하기 때문에 사실적인 높은 품질의 영상을 얻을 수 있다. (그림 2 의 (b))

마지막으로 퓨전 신경망은 회귀 기반 신경망 복원 영상 (x_R)과 인버전 된 영상 (x_G)를 입력으로 받는다. 퓨전 신경망은 회귀 기반 신경망 복원 영상 (x_R)로부터 영상의 구조에 대한 정보를 얻고 인버전 된 영상 (x_G)를 활용하여 영상의 품질을 높인다. 최종적으로 정확하고 높은 품질의 최종 복원 영상 ($\hat{x}$)을 생성한다. (그림 2 의 (c))

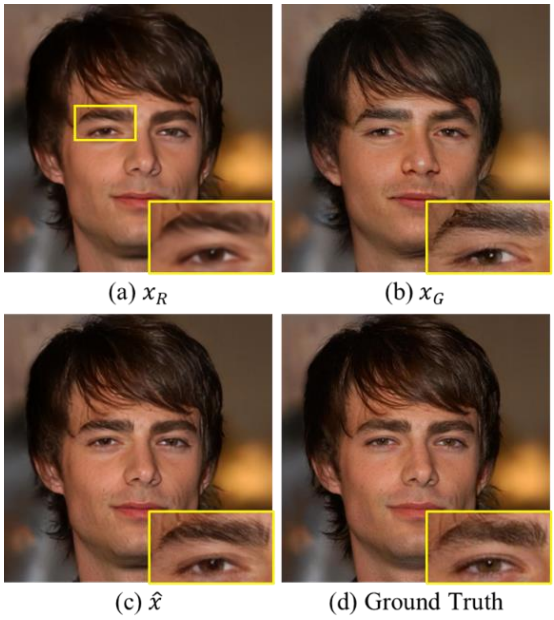

그림 2. 각 생성망을 통해 얻어진 결과물

2.2 적대적 생성망의 활용

기존의 회귀 신경망을 거쳐 얻은 복원 영상을 적대적 생성망의 잠재 공간으로 사상하여 생성자 (G)를 통해 정답 영상과 유사한 높은 품질의 인버전 영상 (x_G)를 얻는다. 이를 위해 영상을 입력 받아 생성망의 잠재 공간으로 사상하는 인코더 (E)를 그림 3 과 같이 설계했다.

인코더는 영상을 입력으로 받아 생성자가 유사한 영상을 생성하도록 16×16 크기의 중간 특징 맵 ($f_{16\times16}$)과 잠재 벡터 (w)들을 추정한다. 중간 특징 맵을 추정하는 인코더는 BDInvert [3]의 인코더 구조와 유사하게 설계했다. BDInvert 는 중간 특징 맵은 인코더로 추정하지만 나머지 잠재 벡터는 최적화를 이용해 찾는다. 이는 상당한 계산 량과 시간을 필요로 한다. 본 연구에서는 잠재 벡터 (w)를 추정하는 네트워크를 추가하여 빠른 계산이 가능하게 했다. 구체적으로는 pSp[8]에서 제안된 map2style 신경망을 사용했다. 각 map2style 신경망은 중간 특징 맵을 추정하는 인코더의 32×32 크기의 중간 특징 맵 ($f_{32\times32}$)를 입력 받아 잠재 벡터를 추정한다.

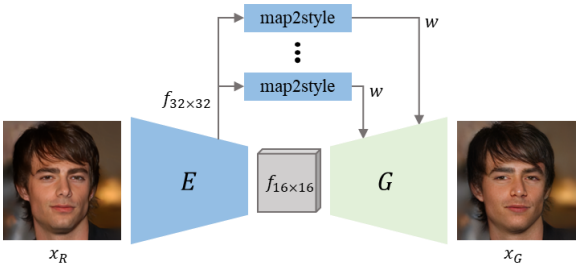

그림 3. 인버전을 위한 인코더 구조

2.3 퓨전 신경망

퓨전 신경망은 회귀 신경망으로 복원된 영상 (x_R)과 적대적 생성망을 통해 인버전 된 영상 (x_G)으로부터의 정보를 조합하여 정확하게 구조를 복원하는 높은 품질의 영상을 생성한다. 퓨전 신경망은 수식 1과 같이 구성하였다. 수식 1에서 $Conv$는 3×3 크기 커널의 합성 곱 층이고 Res는 8 개의 잔차 블록으로 구성된 신경망이다.

$$\hat{x} = Conv(Res(Conv(x_R) + Conv(x_G))) \quad (1)$$

2.4 학습

본 연구에서 제안한 구조는 두단계로 학습된다. 첫 번째 단계에서는 적대적 생성망 활용을 위해 인버전 인코더, 생성자, 구분자를 학습한다. 이 때 미리 학습된 (pretrained) 적대적 생성망의 생성자, 구분자를 사용함으로써 적대적 생성망의 생성 능력을 선행 지식으로 활용한다. 두 번째 단계에서는 퓨전 신경망을 학습한다.

$$L_E = L_1 + \lambda_{per} L_{per} + \lambda_{adv} L_{adv} \quad (2)$$
$$L_F = L_1 + \lambda_{per} L_{per} + \lambda_{fc} L_{fc} \quad (3)$$
$$L_{fc} = \lambda_{cx_R} L_{cx}(x_R, \hat{x}) + \lambda_{cx_G} L_{cx}(x_G, \hat{x}) \quad (4)$$

수식 2 는 인버전 인코더 (E)와 생성자 (G) 학습에 사용된 손실 함수이다. 인버전을 통해 생성된 영상이 정답 영상의 구조를 따르도록 평균 절대 오차 (L_1), LPIPS [12] 오차 (L_{per})가 사용되었다. 인버전을 통해 생성된 영상이 실제 영상과 비슷한 품질을 가지도록 StyleGAN2 [6]에서 제안한 non-saturating GAN loss (L_{adv})를 사용하였다. 구분자는 StyleGAN2 [6]에서 제안한 logistic loss 를 사용하여 인코더, 생성자와 적대적으로 학습시켰다.

수식 3 은 퓨전 신경망 학습에 사용된 손실 함수이다. 퓨전 신경망을 통해 복원된 영상과 정답 영상 간의 평균 절대 오차 (L_1), LPIPS[12] 오차 (L_{per})를 최소화하도록 학습했다. 복원된 영상($\hat{x}$)이 회귀 방법으로 복원된 영상 (x_R)과 인버전 된 영상 (x_G) 모두의 정보를 활용하도록 강제하기 위하여 퓨전 문맥 손실 함수 (L_{fc})를 적용하였다. 퓨전 문맥 손실 함수는 패치단위로 적용되며 수식 4 와 같이 [7]에서 제안한 문맥 손실 함수 (L_{cx})를 이용했다. 문맥 손실 함수는 공간 상의 위치 (spatial)에 관계없이 의미론적으로 (semantic) 가장 가까운 특징 간의 거리를 줄인다.

손실 함수의 계수는 각각 $\lambda_{per} = 10$, $\lambda_{adv} = 0.3$, $\lambda_{fc} = 1, \lambda_{cx_R} = 0.01, \lambda_{cx_G} = 0.05$ 로 설정하였다. 신경망 학습을 위해 Adam 옵티마이저를 사용했다. 첫 번째 단계 학습 시 초기 학습률 (learning rate)는 인코더와 생성자는 10^{-4}, 판별자는 2.5×10^{-5} 이고 총 40,000 번 반복 학습했다. 두 번째 단계 학습 시 퓨전 신경망의 초기 학습률은 10^{-3}이고 8,000 번째 반복 단계에서 학습률을 1/10 로 줄였다. 퓨전 신경망은 40,000 번 반복 학습했고 그 중에 가장 PSNR 이 높은 모델을 사용했다.

3. 실험 결과 및 분석

3.1 실험 세부 사항

본 연구에서 제안한 구조는 다양한 열화 복원 신경망에 적용될 수 있다. 아래의 실험에서는 회귀 신경망으로 [1]에서 제안한 NAFNet 을 사용하여 모션 블러 영상 복원에 대한 결과를 보였다. 학습에는 FFHQ [5] 데이터셋의 512×512 크기의 얼굴 영상을 사용했다. [9]에서 제안된 방법을 따라 71×71 크기의 모션 블러 커널 1,000 개를 생성했다. 각 영상에 무작위로 모션 블러 커널을 적용하여 열화 된 블러 영상과 정답 영상 쌍을 만들었다. 신경망은 블러 영상을 입력으로 받아 정답 영상과 가까운 영상을 복원하도록 학습되었다. 본 연구에서 제안한 방법을 평가하기 위해 학습 데이터를 만든 방법을 CelebA-HQ [4] 3,000 장에 동일하게 적용하여 테스트 데이터 셋을 구성하였다. 기존 회귀 기반 방법 [1,2] 및 생성 기반 방법 [10,11] 과 공평하게 결과를 비교하기 위해 각 방법들을 동일한 학습 데이터셋으로 학습했다.

3.2 정량적 비교

표 1. 이전 방법들과의 정량적 비교

방법	PSNR	SSIM	LPIPS	FID
HINet [2]	29.63	0.81	0.31	40.97
NAFNet [1]	29.53	0.81	0.31	42.18
GFP-GAN [10]	23.09	0.63	0.34	10.73
GPEN [11]	20.60	0.56	0.43	15.48
Ours	28.52	0.76	0.30	9.59

표 1 은 이전 방법들과 제안된 방법의 정량적 비교를 위해 정답 영상과 각 방법이 복원한 영상 간의 PSNR, SSIM, LPIPS, FID 를 측정한 결과이다. 회귀 기반 방법은 파란색, 생성 기반 방법은 초록색, 본 연구에서 제안한 방법은 노란색으로 표시했다. PSNR, SSIM 이 높을수록 정답 영상과 유사한 구조를 가진다. LPIPS 는 정답 영상과의 구조와 품질 차이를 모두 고려하는 점수로 정답 영상과 유사할 수록 점수가 낮다. FID 가 낮을수록 정답 영상과 복원된 영상 사이의 품질 차이가 적다.

본 논문에서 제안한 방법은 이전의 모든 방법보다 좋은 LPIPS, FID 점수를 기록했다. 특히 우리 방법은 회귀 신경망을 사용했지만 적대적 생성망을 활용하여 기존의 회귀 방법보다 현저히 좋은 FID 점수를 보인다. 본 논문에서 제안한 방법은 생성 기

반 방법들보다 월등히 높은 PSNR, SSIM 점수를 기록했다. 이는 블러 제거를 위해 정교하게 고안된 회귀 신경망을 사용하여 다른 생성 기반 방법들보다 정답 영상의 구조를 정확히 복원했기 때문이다.

3.3 정성적 비교

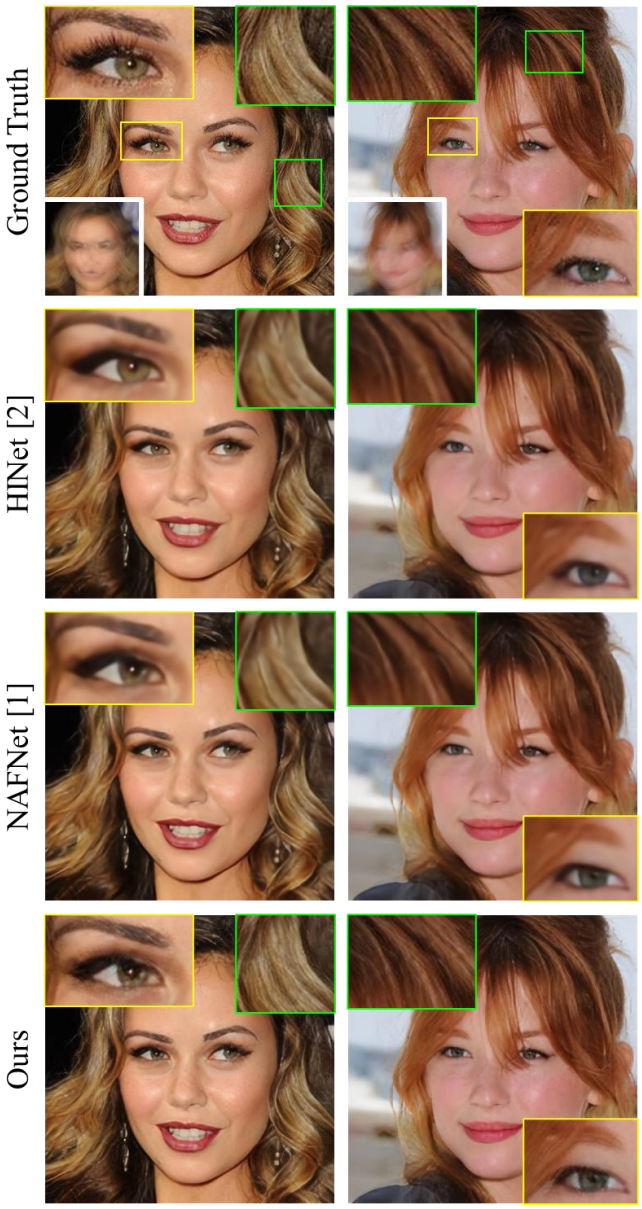

그림 4. 기존 회귀 기반 방법들과의 비교
(정답 영상 왼쪽 하단: 입력 블러 영상)

그림 4 는 기존 회귀 기반 방법들과 본 논문에서 제안한 방법을 정성적으로 비교한 결과이다. 기존 회귀 기반 방법들은 정답 영상의 구조는 비교적 정확히 복원하지만 평균 효과로 인해 다소 낮은 품질의 흐릿한 영상을 생성함을 머리카락, 눈썹 등을 통해 확인할 수 있다. 반면 본 논문에서 제안한 방법은 기존 회귀 기반 방법들과 유사한 정확도를 가지면서도 더 사실적인 좋은 품질의 영상을 생성한다.

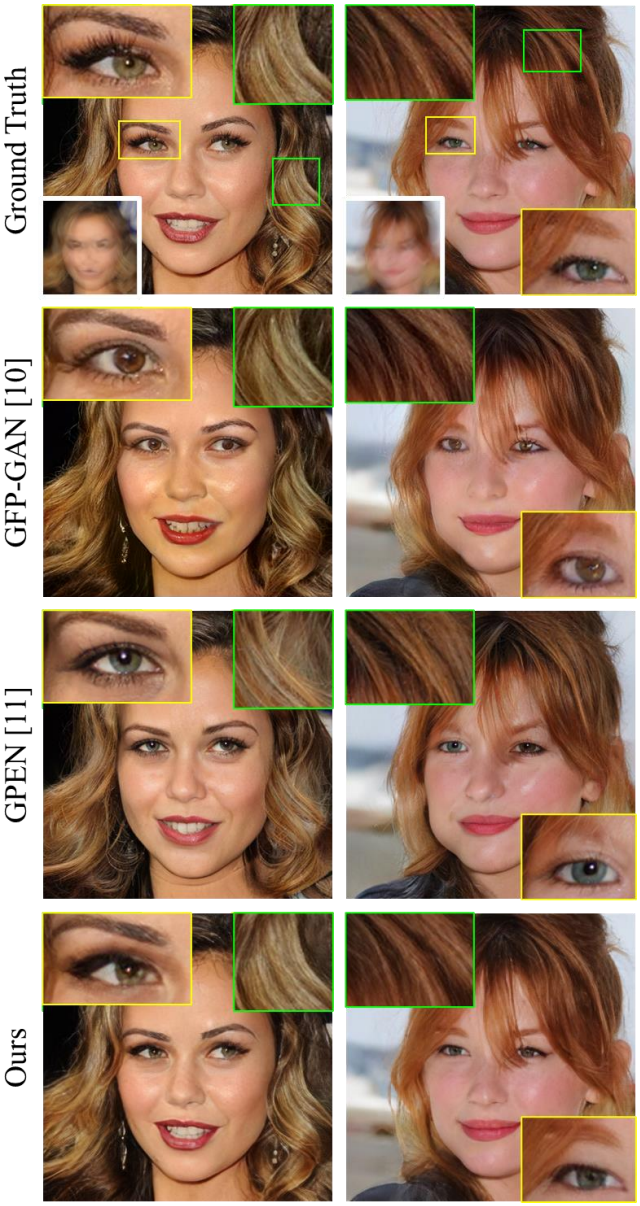

그림 5. 기존 생성 기반 방법들과의 비교
(정답 영상 왼쪽 하단: 입력 블러 영상)

그림 5 는 기존 생성 기반 방법들과 본 논문에서 제안한 방법을 정성적으로 비교한 결과이다. 기존 생성 기반 방법들은 적대적 생성망의 잠재 공간의 제한된 표현력 때문에 정답 영상과 다른 모양의 눈, 입, 치아를 복원함을 확인할 수 있다. 반면 본 연구의 결과는 적대적 생성망을 활용하는 동시에 열화 제거를 위해 정교하게 고안된 회귀 기반 신경망을 사용하기 때문에 정확한 영상 복원이 가능하다.

4. 결론

본 논문에서는 회귀 기반 신경망과 적대적 생성망을 활용하여 정확하고 높은 품질의 영상을 복원하는 방법을 제안했다. 기존 방법과의 정성적, 정량적 비교로 제안된 방법을 이용하여 기존 회귀 기반

방법의 낮은 품질 영상 생성 문제와 기존 생성 기반 방법의 낮은 정확도를 개선할 수 있음을 확인했다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-01906,인공지능대학원지원(포항공과대학교))과 한국연구재단의 지원(No.2020R1C1C1014863)을 받아 수행된 연구임.

참고문헌

[1] Chen, Liangyu, et al. "Simple baselines for image restoration." arXiv preprint arXiv:2204.04676 (2022).
[2] Chen, Liangyu, et al. "HINet: Half instance normalization network for image restoration." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[3] Kang, Kyoungkook, Seongtae Kim, and Sunghyun Cho. "Gan inversion for out-of-range images with geometric transformations." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[4] Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017).
[5] 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.
[6] Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[7] Mechrez, Roey, Itamar Talmi, and Lihi Zelnik-Manor. "The contextual loss for image transformation with non-aligned data." Proceedings of the European conference on computer vision (ECCV). 2018.
[8] Richardson, Elad, et al. "Encoding in style: a stylegan encoder for image-to-image translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[9] Rim, Jaesung, et al. "Real-world blur dataset for learning and benchmarking deblurring algorithms." European Conference on Computer Vision. Springer, Cham, 2020.
[10] Wang, Xintao, et al. "Towards real-world blind face restoration with generative facial prior." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[11] Yang, Tao, et al. "Gan prior embedded network for blind face restoration in the wild." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[12] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.