Journal of Korea Multimedia Society Vol. 22, No. 11, November 2019(pp. 1242-1250)
https://doi.org/10.9717/kmms.2019.22.11.1242

범용 데이터 셋과 얼굴 데이터 셋에 대한 초해상도 융합 기법

문준원†, 김재석††

Super Resolution Fusion Scheme for General- and Face Dataset

Jun Won Mun†, Jae Seok Kim††

ABSTRACT

Super resolution technique aims to convert a low-resolution image with coarse details to a corresponding high-resolution image with refined details. In the past decades, the performance is greatly improved due to progress of deep learning models. However, universal solution for various objects is a still challenging issue. We observe that learning super resolution with a general dataset has poor performance on faces. In this paper, we propose a super resolution fusion scheme that works well for both general- and face datasets to achieve more universal solution. In addition, object-specific feature extractor is employed for better reconstruction performance. In our experiments, we compare our fusion image and super-resolved images from one- of the state-of-the-art deep learning models trained with DIV2K and FFHQ datasets. Quantitative and qualitative evaluates show that our fusion scheme successfully works well for both datasets. We expect our fusion scheme to be effective on other objects with poor performance and this will lead to universal solutions.

Key words: Super Resolution, Fusion Scheme, Deep Neural Network

1. 서 론

이미지 초해상도(super-resolution, SR) 기법은 저해상도(low resolution, LR) 이미지를 고해상도(high resolution, HR) 이미지로 재건하는 영상 처리 기법으로 물체 검출, 감시 카메라에서의 얼굴 인식, 의료 영상처리, 원격 감지, 천문학 영상 등 다양한 분야에서 사용된다 [1]. SR 기법은 고전적인 영상처리 기법으로 일대다(one to many) 사상 문제이기 때문에 정확한 답을 구할 수 없어 지금까지도 해결하기 어려운 문제로 여겨진다. 최근에는 딥러닝 기술의 발전 덕분에 SR 기법의 성능이 매우 크게 증가하였고 이에 따라 최근 십 여년 동안 활발하게 관련 연구가 진행되었으며 특히 단일 이미지를 이용한 SR(single image super-resolution, SISR) 기법이 주목받고 있다 [2].

초해상도 합성 신경구조망(super-resolution convolutional neural networks, SRCNN)[3] 기법은 딥러닝 기법의 대표적인 초기 모델로써, 딥러닝 이전에 사용하였던 드문 부호화(sparse coding) 기반의 방법을 모방하였다. 총 3개의 layer로 이루어진 비교적 얕은 신경망 구조를 가지며, 신경망 모델을 사용한 SR 모델이 고전적으로 사용하던 방법들보다 높은 최

※ Corresponding Author : Jaeseok Kim, Address: (03722) 50, Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea, TEL : +82-0-8876-4018, E-mail : jaekim@yonsei.ac.kr
Receipt date : Aug. 13, 2019, Revision date : Oct. 11, 2019
Approval date : Oct. 28, 2019
† Dept. of Electrical and Electronic Eng., Graduate School, Yonsei University
†† Dept. of Electrical and Electronic Eng., Graduate School, Yonsei University (E-mail : jaekim@yonsei.ac.kr)
※ This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (10080619), and Graduate School of YONSEI University Research Scholarship Grants in 2019.

대 신호 대 잡음비(Peak signal-noise-ratio, PSNR) 값을 얻은 것을 확인하였다. 이 후 깊은 신경망을 효율적으로 학습하는 방법들이 연구되면서 더 큰 신경망을 이용해 성능을 올리는 연구가 진행되었다. 대표적으로 매우 깊은 초해상도 기법(very deep super-resolution, VDSR)[4], 강화된 깊은 초해상도 기법(enhanced deep super-resolution, EDSR)[5] 연구들에서는 보다 깊은 신경망을 구성하여 얕은 신경망 구조보다 높은 성능을 달성하였다. 다른 방향으로는 적대적 생성 신경망(generative adversarial networks, GAN)[6]을 이용한 연구들이 진행이 되었는데 PSNR 지표만을 손실 함수로 사용하였을 때 영상이 흐릿해지는 단점이 관찰되었기 때문이다. 대표적인 연구로 SRGAN(super-resolution generative adversarial networks)[7], 강화된 SRGAN(enhanced super-resolution generative adversarial networks, ESRGAN)[8] 연구들이 있으며 PSNR 기반 손실 함수와 더불어 적대적 손실(adversarial loss) 함수와 인지적 손실(perceptual loss) 함수를 추가하여 현실적이고 자연스러운 질감(texture)으로 복원시켰다[6,7].

SR 기법은 다양한 네트워크 구조, 효과적인 훈련 기법, 효과적인 손실함수의 발달로 인해 점점 좋은 성능을 보이고 있다. 하지만 SR은 근본적으로 일대다 대응 문제로써 역으로 해를 구하기 어려운 문제(ill-posed inverse problem)이기 때문에 여전히 만능(universal)인 SR 모델을 찾는 것은 어렵다. 또한 딥러닝 기법은 데이터 기반으로 네트워크를 학습시키기 때문에 데이터의 분포가 고르지 않다면 데이터 분포의 중심에서 벗어난 물체에 대한 성능이 좋지 않다. 특히 사람의 얼굴에 대한 경우에 성능이 좋지 않았는데, 이를 동기로 본 논문에서는 범용 데이터 셋과 얼굴 데이터 셋에 대한 SR 융합 기법을 제안한다. 먼저 입력 이미지에 대해 얼굴을 검출하고 얼굴이 있는 부분에 대해서 마스크를 구한다. 이를 바탕으로 범용 데이터 셋으로 훈련된 네트워크와 얼굴 데이터 셋으로 훈련된 네트워크를 각각 통과시켜서 얻은 결과 영상들을 융합시킨다. 이 때, 얼굴 데이터 셋에 대한 SR 네트워크 훈련 과정에서 얼굴의 특징을 반영하는 특징 검출기(feature extractor)를 이용하여 얼굴 부분의 SR 성능을 증가시켰다. 본 논문에서 제안하는 융합 기술의 장점은 단순히 얼굴뿐만 아니라 일반적인 이미지 데이터의 분포에서 먼 물체에 대해서 파악이 되면 그것들 역시 융합을 통해 성능을 향상시킬 수 있다는 것이다. 다시 말하면 일반적이지 않은 특징을 가진 물체들에 대해서 모두 좋은 성능을 가지는 범용적인 SR 방법을 얻을 수 있다는 점이다. 두 번째로는 단일의 신경망을 사용하였을 때는 적용하기 힘든 물체 특화된 특징 검출기를 사용할 수 있어서 해당 물체에 대해서 더 높은 복원 성능을 얻을 수 있다는 점이다.

본 논문의 구성은 다음과 같다. 2장에서는 융합을 위해 사용되는 최신의 SR 방법인 ESRGAN과 신경망을 이용한 SR 기법의 단점에 대해서 서술하고, 3장에서는 제안하는 융합 알고리듬을 자세히 기술한다. 그리고 4장에서는 제안하는 알고리듬에 대한 다양한 실험결과와 기존 알고리듬들과의 객관적 및 주관적 비교를 통해 성능을 비교하고 5장에서 결론을 맺는다.

2. ESRGAN 소개 및 문제점

2.1 ESRGAN 모델

SRGAN 네트워크 모델에서는 기존의 PSNR 기반의 학습이 자연스러운 질감을 생성하지 못하는 문제를 해결하기 위해서 적대적 손실 함수와 인지적 손실 함수를 추가적으로 사용한다. 이를 통해 인지적 성능(perceptual quality)를 크게 증대시켰지만 세세한 부분에서 질감이 인위적 결과(artifact)가 관찰되었다. 이를 개선하고 더 나은 인지적 성능을 얻기 위해서 SRGAN을 기반으로 하는 ESRGAN 네트워크 모델이 제안되었다. 첫 번째로 Residual-in-Residual Dense Block (RRDB)이라는 고용량(high capacity)이면서 학습이 효율적인 네트워크 구조를 사용하였고, 두 번째로 진짜인지 가짜인지를 구별하는 기존 GAN을 대체하여 상대적으로 어느 것이 더 현실적인지를 판단하는 Relativistic average GAN(RaGAN)[9]을 사용하였으며, 마지막으로 인지적 손실 함수를 사용할 때 활성화 함수를 통과하기 이전의 값을 사용하여 보다 날카로운 경계를 가지는 영상을 생성하였다. ESRGAN의 신경망 구조는 Fig. 1와 같다.

2.2 신경망을 이용한 SR 기법의 문제점

신경망을 이용한 SR 기법들이 그렇지 않는 고전적 방법들에 비해서 훨씬 좋은 인지적 성능과 PSNR

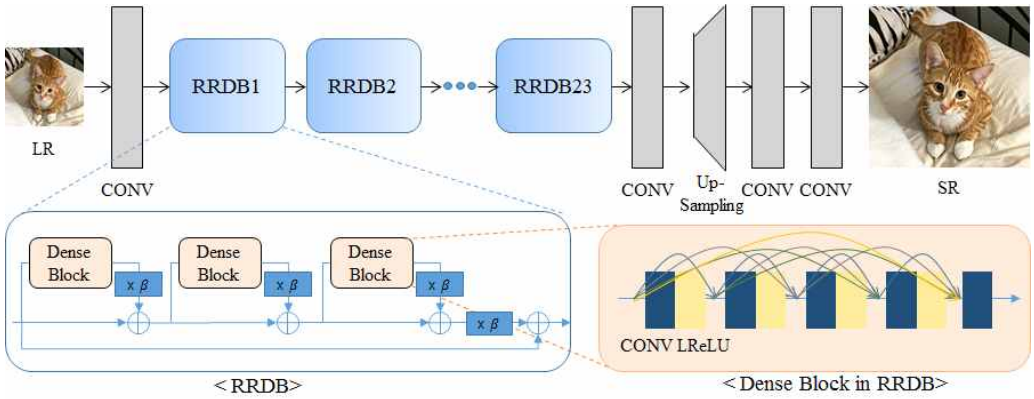

Fig. 1. Network structure of ESRGAN [8].

성능을 보이지만, 사전 기반으로 학습되기 때문에 성능이 데이터 셋의 구성에 의존적이다. 실제에서 만날 수 있는 모든 경우에 대한 데이터를 가지고 훈련시키는 것은 데이터 수집, 훈련할 데이터의 양, 그것을 수용할 수 있는 네트워크의 용량 등을 고려했을 때 불가능에 가깝기 때문에 현실적으로 모든 경우에 SR이 좋은 성능을 내게 하는 것은 매우 어렵다. 다시 말해, 데이터 분포가 밀집해 있다면 그 부분에 대한 성능은 제한된 크기의 단일 네트워크로 좋은 성능을 보일 수 있지만, 밀집한 영역에서 벗어난 물체의 경우는 SR의 성능이 좋지 않다. Fig. 2의 경우는 사람의 얼굴에 대한 SRGAN과 ESRGAN 신경망들의 결과들로, 사람의 얼굴이 가지는 중요한 특징인 눈이나 이 부분이 흐릿하게 생성되는 문제가 있다. 따라서 이러한 특수한 경우에 대해서 SR 성능을 높일 수 있는 방법이 요구된다(차이를 명확하게 보기 위해서는 전자 문서에서 그림을 확대하여 보는 것을 권장한다).

3. 제안한 방법

3.1 제안한 알고리듬의 개요

본 논문에서 우리는 제한된 신경망 구조를 이용하여 범용적으로 좋은 성능을 가지기 위한 SR 융합 기법을 제안한다. 먼저 다양한 물체들로 이루어진 이미지들을 데이터 셋으로 훈련시켰을 때 얻어지는 SR 네트워크를 이용해 SR 성능이 낮은 물체 목록을 작성한다. 그 후, 해당 목록에 대해서 물체 검출 알고리즘을 통해 검출 지도를 얻는다. 대표적인 물체 검출 알고리즘으로 Mask R-CNN[11]이 있으며, 그 외에도 특정 물체에 대한 검출 알고리즘이 다양하게 연구되고 있다. 검출 지도가 구해지면 서로 다르게 학습된 SR 네트워크를 통해 SR 결과를 각각 생성한다. 이 때, 다양한 물체로 이루어진 데이터 셋으로 훈련된 SR 네트워크는 검출 지도에서 물체가 차지하지 않는 배경(background)에 적용되게 된다. 최종적으로 다른 신경망에서 얻어진 SR 이미지들을 자연스럽게 융합시키는 과정을 거친 후 융합 SR 이미지를 생성한다. 제안하는 알고리듬의 개요도는 Fig. 3과 같다.

3.2 얼굴과 얼굴이 아닌 부분에 대한 SR 융합 기법

본 논문에서는 물체 목록이 얼굴 하나인 경우에

Fig. 2. SR results for various SR algorithms. The input image is 04723 from [10]. Scale factor is x4 and the size of LR image is 36×36.

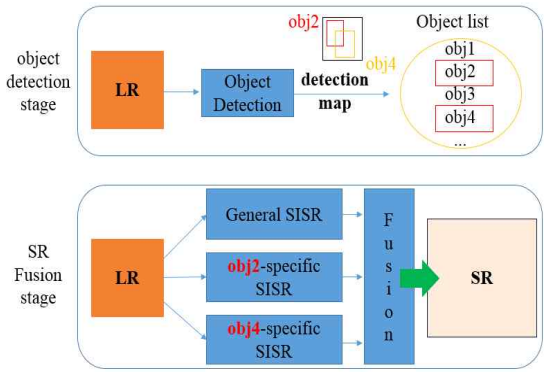

Fig. 3. The block diagram of proposed scheme.

대해서 분석하였다. 컴퓨터 비전 분야에서는 사람 얼굴에 대해 검출, 분류, 동일인물 판별, 연령 판별 등 다양한 과제들로 연구되고 있다. 본 논문에서는 얼굴 검출을 위해서 yoloface 알고리즘[12]을 사용하였으며, 얼굴에 특화된 네트워크를 학습하기 위해 Flickr-Faces-HQ(FFHQ) 데이터 셋[10]을 사용하였고, VGGFace2 데이터 셋[13]을 이용해 다양한 과제로 훈련된 신경망 구조를 특징 검출기로 사용하였다. 반면, 얼굴이 아닌 부분에 대해서는 기존 ESRGAN 알고리즘과 마찬가지로 DIV2K 데이터 셋[14]을 사용하였고, 훈련 방법은 동일하게 사용하였다.

얼굴에 특화된 SR 신경망으로부터 생성된 SR 이미지를 SR_{Face}, 보편적인 이미지로 훈련된 SR 신경망으로부터 생성된 SR 이미지를 $SR_{General}$이라고 할 때, SR_{Face}은 얼굴인 부분에서 성능이 좋고 $SR_{General}$은 얼굴이 아닌 배경에서 성능이 좋다. 따라서 검출 지도를 이용해 각각이 장점인 영역에 적응적으로 SR 네트워크를 적용시킨 결과 이미지들을 융합하면 단일의 네트워크를 적용시킨 것보다 높은 성능을 얻을 수 있다. 이를 위해 다음과 같이 SR_{Face}와 $SR_{General}$를 가중합(weighted sum)하여 융합 이미지 SR_{Fusion}를 생성한다.

$$SR_{Fusion} = w_{General}SR_{General} + w_{Face}SR_{Face} \tag{1}$$

이 때, 얼굴 마스크의 경계 부분에서 부자연스럽지 않도록 하기 위해 w_{Face}값은 얼굴 검출 마스크를 메워넣기(padding) 한 후 가우시안 필터를 합성곱(convolution)하여 평활화 시킨다. 메워넣기를 하는 것은 얼굴 검출에 사용한 알고리즘이 검출하는 지도의 크기가 작게 나오는 경우가 많기 때문이고, 만약 검출 알고리즘이 물체를 크게 마스크로 잡는다면 생략해도 될 것이다. 합성곱을 하는 단계에서는 얼굴 검출 마스크의 너비와 높이를 각각 w, h라고 할 때, 가우시안 필터 F 를 다음과 같이 구한다.

$$F(i,j) = \frac{1}{Z}exp(-\frac{(i-(w+1))^2}{2\sigma_x^2})\exp(-\frac{(j-(h+1))^2}{2\sigma_y^2}),$$
$$i \in [1, 2w+1],\ j \in [1, 2h+1] \tag{2}$$

여기서 Z는 F의 원소 합이 1이 되도록 하는 정규화 상수이며, σ_x^2, σ_y^2는 가우시안 함수의 분산으로, 본 논문에서는 각각 $(0.8w)^2, (0.8h)^2$ 값을 사용하였다. w_{Face} 값이 결정되면 배경 부분의 가중치 $w_{General}$는 아래와 같이 구해진다.

$$w_{General}(i,j) = 1 - w_{Face}(i,j) \tag{3}$$

최종적인 융합 과정에 대한 도식도는 Fig. 4와 같다.

3.3 얼굴 특징 손실 함수

제안하는 융합 알고리듬에서는 기존 SR 신경망에서 사용하는 다양한 손실함수와 함께 특정 물체에 특화된 특징 검출기(object-specific feature extractor)의 값을 추가로 사용하였다. 이는 물체에 따라 신경망을 학습시키기 때문에 가능한 손실 함수이다. 특징 검출기는 다른 과제(task)에서 학습된 신경망 구조의 일부를 의미하며, 다른 과제에서 충분히 검증되고 적용분야의 교집합이 많을 경우 원하는 과제에서 보다 좋은 성능을 가지도록 학습하게 된다. 예로써, 얼굴에 대해서 특징 검출기를 만들기 위해서는 얼굴 인식, 얼굴 감지, 얼굴 증명 등과 같은 과제를 위한 신경망을 학습시키고 그 신경망의 일부를 떼어내면 된다. 얼굴의 경우 이미 해당 분야가 활발히 진행되어서 온라인으로 신경망 구조와 학습된 가중치가 공유되어 있다[13]. 우리는 이 중에서 Senet50_ft_dag 신경망 구조를 사용하였고, pool5_7×7_s1 층의 결과를 특징 추출기 값으로 사용하였다. 추출된 값은

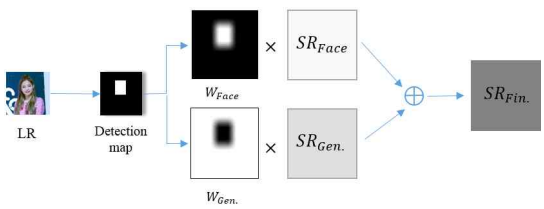

Fig. 4. The proposed SR fusion scheme.

2048×2×2 값을 가지게 되며 2×2 배열의 경우는 *l2*-norm으로 거리를 측정하고 그것의 결과인 2048개의 원소로 이루어진 벡터는 얼굴 서술자(descriptor)로써 역할을 하기 때문에 코사인 유사도(cosine similarity)를 통해 생성된 SR 이미지와 기준 HR 이미지와의 유사도를 측정하였다. 따라서 제안하는 알고리듬에서는 손실함수로써, PSNR 기반의 손실(L_{PSNR}), 인지적 손실(L_{Percep}), 적대적 손실(L_{Gan}), 그리고 얼굴 특징에 특화된 인지적 손실($L_{VGGFace}$)로 아래와 같다.

$$D_k(x) = \sqrt{d_{k,1,1}^2(x') + d_{k,1,2}^2(x') + d_{k,2,1}^2(x') + d_{k,2,2}^2(x')} \quad (4)$$

$$L_{VGGFace} = 1 - (\frac{D(SR) \cdot D(HR)}{|D(SR)\|D(HR)|})^n \quad (5)$$

$$L_{tot} = \lambda_{psnr} L_{PSNR} + \lambda_{per} L_{percep} + \lambda_{gan} L_{Gan} + \lambda_{face} L_{VGGFace} \quad (6)$$

여기서 x'은 이미지 x를 입력으로 받아 224×224 크기로 선형 보간된 결과이며 d는 특징 추출기를 통과시킨 2048×2×2 크기의 결과 행렬이다. 그리고 2048 크기의 압축된 특징 서술자 D의 코사인유사도를 손실함수로 적용시키기 위해서 (6)과 같이 1의 보수 값을 사용했으며 학습을 원활히 하기 위해서 유사도에 지수 승을 하였다. n이 너무 크면 수렴하지 않고 작으면 학습이 힘들기 때문에 실험적으로 6을 사용하였다.

4. 실험 결과 및 고찰

4.1 훈련 세부사항 및 평가 지표 소개

본 논문에서 진행한 실험들은 초해상도 복원의 스케일 배율(scale factor)은 x4을 사용하였으며 네트워크를 학습하기 위해서 그래픽 카드 RTX2080을 사용하였고 다운 샘플링 및 이미지 패치 자르기(crop)는 Matlab 툴의 imresize 함수를 사용하였다. 실험에서 사용한 데이터 셋은 DIV2K[14]와 FFHQ[10]이다. DIV2K는 다양한 장면으로 구성된 HD(high definition)급의 고품질 데이터 셋이며 훈련 데이터 셋과 유효성 검사 데이터 셋으로 각각 800개와 100개로 이루어져 있다. 훈련을 용이하게 하기 위해 120x120 패치로 잘라서 사용하였고 미니 배치의 크기는 16이며 반복(iteration) 횟수는 200,000번 학습하였다. GAN 네트워크와 특징 추출기, 손실 함수 및 가중치는 ESRGAN[8]과 동일하게 하였다. FFHQ 데이터 셋은 1024×1024 고품질 사람 얼굴 이미지들로 이루어졌으며 다양한 연령, 민족, 배경, 밝기 등을 포함되어 있다. 훈련 데이터 셋은 68,779개, 유효성 검사 데이터 셋은 1200개의 이미지로 나누었다. 또한 작은 이미지에서의 초해상도 복원이 더 힘들기 때문에 어려운 경우에도 복원을 잘 하기 위해 1024×1024에서 80×80 ~ 148×148로 랜덤 다운샘플링 한 것을 사용하였다. 그 후, 배경을 제외하고 오로지 얼굴에 대해서만 훈련하고 시험하기 위해서 yoloface[12]라는 얼굴 검출 알고리즘을 통해 추출하였다. 이 때 다운샘플링을 하는 것은 해상도가 낮을수록 얼굴에 대한 초해상도 복원의 결과 성능이 현저히 저하되기 때문에 이러한 경우에서도 복원이 잘되도록 하기 위해서이다. 미니 배치 크기는 8로 줄여서 사용하였는데 이는 얼굴 특징 추출기가 추가되어 그래픽 카드의 메모리가 부족해졌기 때문이다. 다른 훈련 과정은 모두 DIV2K 데이터 셋을 훈련 할 때와 같으며, 손실함수에서 $L_{VGGFace}$를 사용할 때는 가중치 λ_{face} 값을 10으로 사용하였다.

평가 지표의 경우 전통적으로 많이 쓰였던 PSNR과 structural similarity(SSIM)을 사용하였다. PSNR과 SSIM은 때때로 사람이 평가하는 성능과 거리감이 있다고 많이 알려지고 있기 때문에 정량적 평가뿐만 아니라 정성적 평가도 함께 진행하였다.

4.2 데이터 셋에 따른 결과와 제안하는 융합 알고리듬의 결과에 대한 정성적 평가

제안하는 융합 알고리듬은 네트워크의 구조나 훈련 방법을 개선시키는 것이 아니기 때문에 다른 SR 알고리듬들과의 비교보다는 네트워크 구조와 훈련 방법을 ESRGAN 알고리듬으로 고정시킨 상태에서 데이터 셋에 따라 신경망을 이용한 SR 방법의 성능을 분석하였다. 추가로 제안하는 융합 방식에서는 얼굴 특징기를 추가로 손실함수로 사용하였기 때문에 얼굴 특징기를 추가했을 때의 성능 개선을 살펴보았다. 우리가 원하고자 하는 융합 알고리듬은 범용적으로 성능이 좋게 만드는 것이며 예시로써 얼굴과 비-얼굴 (non-face) 부분에 대해 나누어 융합을 한다. 먼저 3가지 이미지들에 대해서 각각 다른 데이터 셋으로 훈련시켰을 때와 제안하는 특징기를 사용할 때 및 융합 알고리듬을 사용할 때의 결과들을 정성적인 평가를 진행하였고, 그 후에 DIV2K 데이터 셋과 FFHQ 데이터 셋에 대해 정량적인 평가 지표들을 비

교하였다.

아래의 Fig. 5에서는 얼굴 이미지에 대해 데이터 셋과 특징 추출기에 따른 성능 차이를 살펴본다. Fig. 5(a)는 LR 이미지를 선형 보간한 이미지이며 Fig. 5(b)는 DIV2K 데이터 셋을, Fig. 5(c)는 FFHQ 데이터 셋을, Fig. 5(d)는 DIV2K와 FFHQ 데이터 셋을 함께 이용하여 훈련시킨 신경망의 결과 SR 이미지들이다. 먼저 Fig. 5(b) 결과를 보면 입력 이미지에 비해 더 선명해지지만 눈썹과 눈의 구분이 명확하지 않고 콧구멍과 콧볼이 복원되지 않았으며 입술 역시 일그러져있다. 귀걸이 부분은 가장 명확하게 보이는 것을 확인할 수 있다. Fig. 5(c)는 얼굴 데이터 셋을 이용하였기 때문에 왼쪽 콧구멍과 입술 부분은 자연스럽게 복원되었다. 하지만 오른쪽 콧구멍과 콧볼, 그리고 눈 부분은 여전히 좋지 않았다. Fig. 5(d)의 경우는 Fig. 5(c)와 전체적으로 비슷하며 귀걸이 부분은 Fig. 5(c)보다 선명하고 Fig. 5(b)보다는 덜 선명하다. Fig. 5(e)는 얼굴 특징 검출기를 이용한 손실 $L_{VGGFace}$을 포함하여 학습되었으며 데이터 셋은 FFHQ를 사용한 결과이다. 얼굴 특징 검출기를 사용하지 않는 Fig. 5(c)와 비교해보았을 때, 콧구멍과 콧볼이 더 자연스럽게 복원되고 눈썹과 눈도 확실하게 구분되는 것을 볼 수 있다. Fig. 5(f)는 Fig. 5(h)를 w_{Face} 값으로 Fig. 5(b)와 Fig. 5(e)를 융합시킨 결과이다. 마스크의 값이 얼굴 부분을 제대로 검출하였기 때문에 Fig. 5(e)와 마찬가지로 얼굴을 잘 복원하였으며, 마스크의 경계부분에서도 자연스러운 것을 볼 수 있다. Fig. 5(g)는 HR 이미지를 나타낸다.

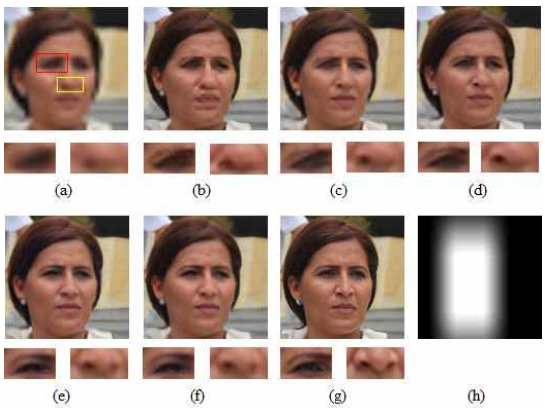

Fig. 5. (a) Input(bicubic interpolated) (b) DIV2K (c) FFHQ (d) DIV2K+FFHQ (e) FFHQ with $L_{VGGFace}$(proposed) (f) Fusion result(proposed) (g) HR image (h) w_{Face}.

Fig. 6의 경우 빨간색 상자가 얼굴 중 코 부분이고 노란색 상자가 얼굴이 아닌 타일 영역이다. 코 부분의 경우, $L_{VGGFace}$를 사용하지 않은 Fig. 6(b)-(d) 부분에서 모두 자연스럽지 않게 복원되었으며 $L_{VGGFace}$를 사용한 상대적으로 Fig. 6(e)는 코의 형태를 가지도록 복원되었다. 얼굴이 아닌 부분인 노란 색 상자 부분은 Fig. 6(b)가 가장 선명했으며 Fig. 6(d)가 다음으로 선명하고 Fig. 6(c)와 (e)의 경우는 선이 두꺼워지고 깔끔하지 않게 복원이 되었다. 얼굴 부분에는 질감이 강한 부분이 상대적으로 적기 때문에 날카롭게 복원이 되지 않는 것으로 추측된다. Fig. 6(f)의 융합 결과 이미지에서는 코와 타일 부분에서 가장 성능이 좋았던 Fig. 6(e)와 Fig. 6(b)를 함께 가져오는 결과를 보인다. 이 외에 눈썹과 이가 제안하는 Fig. 6(f)에서 가장 자연스럽게 복원되었다.

정리하면, DIV2K로 학습된 것은 얼굴 데이터 셋에서 실제 얼굴의 특징을 제대로 반영하지 못하지만 얼굴이 아닌 부분에서는 비교적 선명하게 복원하였으며, FFHQ를 사용한 데이터 셋에서는 역으로 얼굴 부분을 보다 잘 복원하게 된다. 데이터 셋을 섞어 훈련시킨 경우 얼굴 데이터 셋이 비교적 많이 포함되기 때문에 얼굴 부분을 DIV2K만을 이용한 것보다는 잘 복원시키지만, Fig. 6(d)와 같이 얼굴의 특징을 반영하지 못하는 경우가 종종 발생한다. 보다 얼굴을 잘 복원하기 위해서 얼굴 특징 검출기를 이용해 $L_{VGGFace}$

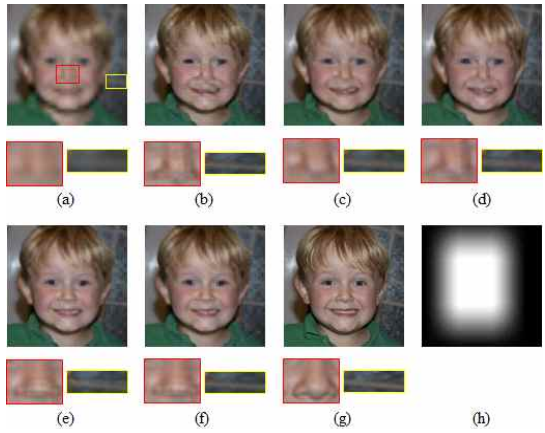

Fig. 6. (a) Input(bicubic interpolated) (b) DIV2K (c) FFHQ (d) DIV2K+FFHQ (e) FFHQ with $L_{VGGFace}$(proposed) (f) Fusion result(proposed) (g) HR image (h) w_{Face}.

를 학습에 사용하게 되면 얼굴 특징에 따라 복원을 하려고 하기 때문에 콧볼, 눈, 이 등 이목구비의 특징을 잘 살리는 경향성이 있다. 최종적으로 제안하는 융합 알고리듬은 데이터 셋에 따른 장단점과 물체에 특화된 특징 검출기를 모두 이용하기 때문에 경우에 따라 범용적으로 성능이 좋게 초해상도 복원이 되는 것을 볼 수 있다.

정량적인 평가를 위해서 4.1에서 설명한 PSNR, SSIM을 사용하였고 대상으로 DIV2K 이미지 데이터 셋과 FFHQ 데이터 셋을 사용하였다. PSNR과 SSIM의 경우는 높은 값을 가질수록 성능이 좋음을 나타낸다. DIV2K 데이터 셋에서는 얼굴을 포함하는 이미지가 5% 이하이며, 그마저도 얼굴이 차지하는 부분은 대부분 10% 미만이다. 각 데이터 셋에 대한 성능 지표 결과는 아래와 같다. 두 데이터 셋에 대해 PSNR의 경우 명확히 훈련된 데이터 셋 종류로 테스트를 했을 때 높은 값을 가지는 것을 볼 수 있고, SSIM의 경우는 데이터 셋과 관계없이 DIV2K 데이터 셋으로 훈련시킬 때 높은 값을 가지는 경향이 있다. 이는 얼굴의 특징이 다른 물체들보다 더 특이하여 정답과 비슷한 구조를 가지는 SR 이미지를 생성하는 것이 어렵기 때문으로 추측된다. DIV2K 이미지를 테스트 했을 때는 얼굴을 거의 포함하지 않기 때문에 DIV2K로 훈련한 것이 DIV2K+FFHQ로 훈련한 것보다 SSIM에서 상대적으로 높은 값을 가졌다. 제안하는 융합 알고리듬의 결과는 PSNR, SSIM 각각 가장 높거나 그 다음으로 높은 값을 보이는 것을 확인할 수 있다. 이는 각 데이터 셋으로 훈련된 결과 중 성능이 좋은 부분을 선택적으로 융합하게 되고, 추가적으로 물체 특징 검출기 이용해 성능 향상을 시킬 수 있기 때문이다.

4.3 데이터 셋에 따른 결과와 제안하는 융합 알고리듬의 결과에 대한 주관적 평가

정량적인 지표인 PSNR과 SSIM의 경우 실제 사람의 주관적인 평가와 괴리감이 있다고 널리 알려져 있다. 따라서 본 논문에서는 Mean Opinion Score 방법을 통해 주관적 화질평가를 진행하였다. 테스트 이미지의 경우 DIV2K와 FFHQ에서 각각 30개씩 추출하여서 1점부터 5점까지 5개의 보기 중에 평가하도록 하였다. 평가할 이미지들은 각 테스트 이미지마다 5개씩 주어지며 각각 Table 1과 같은 훈련과정에 따라 생성되었다. 테스트 이미지 1개당 30초의 시간 내에 평가를 하도록 가이드하였으며 빠른 선택을 돕기 위해 LR 이미지를 1점, HR 이미지를 5점으로 참고하라고 함께 제시하였다. 또한 모든 선택지에 대해 같은 점수를 주지 않고 최대한 차등하여 평가하도록 지시하였다. 실험 대상은 50명이며, 각각 DIV2K 15개와 FFHQ 15개씩 평가하도록 진행하였다. 실험 결과에 대한 요약은 Table 2와 같이 하였으며, 평균과 표준편차 값으로 표시하였다.

DIV2K에 대한 테스트 결과로, DIV2K+FFHQ 데이터로 훈련한 모델이 가장 성능이 좋았는데, 이것은 데이터의 총량이 증가하였기 때문으로 추측된다. DIV2K 데이터로 훈련된 경우는 제안한 합성 결과와 비슷하게 나왔는데, 이는 DIV2K 데이터 셋에 얼굴이 거의 포함되어 있지 않았기 때문이다. FFHQ에 대한 테스트 결과, 제안된 합성 결과와 얼굴 특징 손실함수를 추가한 것이 큰 차이로 높은 점수를 얻었다. DIV2K로 훈련된 경우 가장 성능이 좋지 않았으

Table 1. Quantitative comparison for DIV2K and FFHQ dataset

Test Dataset	Training process	PSNR	SSIM
DIV2K	DIV2K	28.793	0.799
	FFHQ	28.315	0.788
	DIV2K+FFHQ	28.588	0.793
	FFHQ+$L_{VGGFace}$	28.317	0.791
	Fusion	28.846	0.797
FFHQ	DIV2K	28.438	0.845
	FFHQ	28.911	0.842
	DIV2K+FFHQ	28.720	0.847
	FFHQ+$L_{VGGFace}$	29.037	0.841
	Fusion	29.227	0.846

Table 2. Subjective Quality Assessment (average±standard deviation) Results for DIV2K and FFHQ dataset

Training process \ Test image	DIV2K	FFHQ	Total
DIV2K	3.69±0.91	1.95±0.82	2.82±1.23
FFHQ	2.91±0.98	2.38±0.89	2.64±0.97
DIV2K+FFHQ	4.09±0.86	2.34±0.85	3.21±1.22
FFHQ+$L_{VGGFace}$	3.27±0.94	3.55±0.97	3.41±0.96
Fusion	3.68±0.89	3.75±0.91	3.71±0.90

며, FFHQ로 훈련된 것과 DIV2K+FFHQ 데이터로 훈련 된 것은 거의 차이가 없었는데 이를 통해 복원이 잘 되지않는 얼굴의 경우 일반적인 데이터를 추가해도 성능이 크게 나아지질 않는 것을 확인할 수 있었다. 마지막으로 두 데이터 셋 모두에 대한 합계 결과이다. 제안하는 합성 방법을 사용한 것이 평균 점수도 가장 높고 표준 편차는 가장 낮은 것을 확인할 수 있는데, 이는 보편적으로 높은 점수에 집중적으로 평가되었다는 것을 의미한다. 따라서 논문의 목적인 범용적 초해상도 복원을 성공적으로 수행하는 것을 알 수 있다. 또한, 본 논문에서는 DIV2K만을 이용하여 훈련된 신경망으로 $SR_{General}$를 생성하였는데 DIV2K+FFHQ 훈련 데이터 셋을 사용하게 되면 더 높은 성능을 달성할 수 있을 것으로 예상된다.

5. 결 론

본 논문에서는 신경망을 이용한 초해상도 복원 기법이 특정 물체에서는 성능이 현저히 떨어질 수 있다는 사실을 지적하고, 그것을 발견했을 때 해당 물체에도 성능이 좋게 만들어 최종적으로 범용적인 초해상도 복원 알고리듬을 만들 수 있는 융합 기법에 대해 소개하였다. 신경망 알고리듬의 발전으로 매우 큰 성능 향상을 이루었지만, SR 기법의 경우는 정답을 찾는 것이 매우 어렵기 때문에 성능 향상이 가능한 한계(margin)가 아직 남아있다. 많은 논문들에서 네트워크의 구조나 훈련 방법 등을 개선하는 연구를 진행하고 있지만, 그러한 방법들 역시 특정 물체에 대해서는 성능이 떨어질 수 있다는 가능성이 내포되어 있다. 제안하는 융합 기법은 단일 네트워크를 이용한 SR 신경망에 비해 범용성 측면에서 높은 성능 향상을 이룰 것으로 보인다. 추가적으로, 물체마다 다른 네트워크 구조의 SR를 사용하여 결과를 융합하면 더 좋은 성능을 낼 수 있지만, 같은 네트워크를 사용하면 가중치들을 공유하여 간소화될 가능성이 있고 알고리듬 설계 방식에 따라 똑같은 구조로 설계하는 것이 프로세싱 유닛을 줄일 수 있기 때문에 본 논문에서는 같은 네트워크로 실험하였다. 앞으로의 연구에서는 검출 단계의 신경망과 서로 다른 물체에 대한 SR 신경망들이 부분 공유하는 방법으로 처리속도 간소화에 대해 진행할 것이다.

REFERENCE

[1] S. Anwar, S. Khan, and N. Barnes, “A Deep Journey into Super-resolution: A Survey,” *arXiv Preprint arXiv:1904.07523*, 2019.

[2] D.H. Lee, H.S. Lee, K.J. Lee, and H.J. Lee, “Fast Very Deep Convolutional Neural Network with Deconvolution for Super-resolution,” *Journal of Korea Multimedia Society,* Vol. 20, No. 11, pp. 1750-1758, 2017.

[3] C. Dong, C.C. Loy, K. He, and X. Tang, “Learning a Deep Convolutional Network for Image Super-resolution,” *Proceeding of European Conference on Computer Vision,* pp. 184-199, 2014.

[4] J. Kim, J. Lee, and K. Lee, “Accurate Image Super-resolution Using Very Deep Convolutional Networks,” *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1646-1654, 2016.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. Lee, “Enhanced Deep Residual Networks for Single Image Super-resolution,” *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 136-144, 2017.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warder-Farley, S. Ozair, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems,* pp. 2672-2680, 2014.

[7] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, et al., "Photo-realistic Single Image Super-resolution Using a Generative Adversarial Network," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4681-4690, 2017.

[8] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, et al., "Esrgan: Enhanced Super-resolution Generative Adversarial Networks," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 63-79, 2018.

[9] A. Jolicoeur-Martineau, "The Relativistic Discriminator: A Key Element Missing from Standard GAN," *arXiv Preprint arXiv:1807.00734*, 2018.

[10] T Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4401-4410, 2019.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proceeding of the IEEE Conference on Computer Vision,* pp. 2961-2969, 2017.

[12] J. Redmon and A. Farharhadi, "Yolov3: An Incremental Improvement," *arXiv Preprint arXiv:1804.02767*, 2018.

[13] Oxford University, Information Engineering, http://www.robots.ox.ac.uk/~albanie/pytorch-models.html (accessed July 22, 2019)

[14] E. Agustsson and R. Timofte, "Ntire 2017 Challenge on Single Image Super-resolution: Dataset and Study," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 126-135, 2017.

문 준 원

2013년 연세대학교 전기전자공학과 학사 졸업
2013년~현재 연세대학교 전기전자공학과 통합과정
관심분야 : Image processing, Super-resolution, hardware design

김 재 석

1977년 연세대학교 전기전자공학과 학사 졸업
1979년 한국과학기술원 전기 및 전자공학과 석사 졸업
1988년 Rensselaer Polytechnic Institute NY, USA 전기전자공학과 박사
1996년~현재 연세대학교 전기전자공학과 교수
관심분야 : Communication IC design, high performance Digital Signal Processor VLSI design, and CAD S/W