이미지 초해상화 및 인페인팅 합동 학습을 위한 단계적 처리 모델

*손채연, *김수예, **김희권, *김문철
*한국과학기술원, **한국전자통신연구원
thscodus97@kaist.ac.kr, sooyekim@kaist.ac.kr, hkkim79@etri.re.kr, mkimee@kaist.ac.kr

A Step-by-Step Approach for Joint Learning of Image Super-Resolution and Inpainting

*Chaeyeon Son, *Soo Ye Kim, **Hee Kwon Kim, *Munchurl Kim
*Korea Advanced Institute of Science and Technology,
**Electronics and Telecommunications Research Institute

요 약

본 논문에서는 꾸준히 연구되어 오던 이미지 복원 문제에서 초해상화와 인페인팅이라는 복합적 이미지 복원을 동시에 처리하는 해결 방법을 제안한다. 초해상화는 국지적 픽셀 정보를 이용하여 고해상도의 영상을 복원하고, 인페인팅은 이미지 전체 정보를 활용하여 영상 내 비어 있는 영역을 생성해야 하므로, 이러한 두 가지 영상 복원 기법을 동시에 수행하는 것은 상당히 어려운 문제이다. 그렇기에 인페인팅과 초해상화는 이미지 복원에서 널리 활용되는 기술인 만큼 동시에 해결할 수 있는 기법에 대한 수요는 있음에도 지금까지 거의 연구되지 않았다. 본 논문은 초해상화 및 인페인팅 합동 처리에 있어 복합적인 정보를 모두 다뤄야하는 네트워크가 서로의 성능을 저하시키지 않도록 개략적 복원 네트워크 (Coarse network), 디테일 복원 네트워크 (Refinement network), 초해상화 네트워크 (SR network)로 분리하여 초해상화 및 인페인팅 합동 처리를 수행하며, 각 단계마다 결과 영상을 얻어 스케일 별 정답 영상과 손실함수를 계산하여 복합적인 성능을 올릴 수 있는 방법을 제시한다. 또한 순차적 단일 모델에 비하여 인페인팅과 초해상화를 합동 학습하는 제안 모델이 개선된 화질의 결과 영상을 획득할 수 있다는 것을 실험적으로 보인다.

1. 서론

손상된 이미지를 복원하는 문제는 영상 처리분야에서 지금까지도 활발히 연구되고 있는 전통적인 문제이다. 대표적인 영상 복원 기술로는 낮은 해상도를 높은 해상도로 만드는 초해상화, 움직임 등으로 인해 발생한 블러를 제거하는 디블러링, 영상이 손상되어 발생한 노이즈를 제거하는 디노이징, 마스크 처리된 손실된 픽셀 부분을 자연스럽게 채워 넣는 인페인팅 등이 있다. 실제 상황에서 활용되는 경우를 고려하다 보니, 최근에는 단순한 단일 손상보다는 여러 손상이 복합되어 있는 영상을 동시에 해결하려는 연구의 중요성이 점점 커지고 있는 추세이다. 그 중 전체적인 이미지의 퀄리티의 향상을 위해 사용되는 초해상화와 스크래치 등으로 손상된 부분을 복원하거나 사용자가 원하는 특정 부분을 지우기 위해 사용되는 인페인팅은 각각 영상 복원 분야에서 중요성과 활용되는 빈도가 모두 높다. 그러므로 이 둘을 동시에 해결할 수 있는 기법 역시 요구되고 있다.

초해상화는 앞서 언급되었듯이 저화질의 이미지를 고화질의 이미지로 변환시키는 기법이며, 딥러닝 기법이 도입되기 이전에는 다양한 보간법 (Bilinear interpolation, Bicubic interpolation

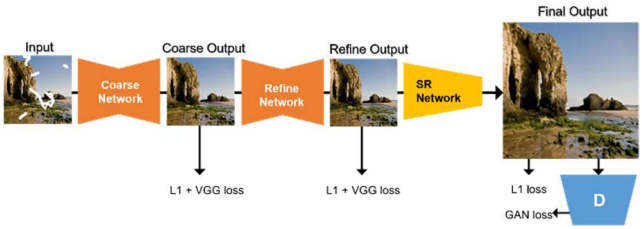

그림 1. 전체 네트워크 모식도와 단계별 손실 함수

등)을 활용하여 해당 문제를 해결했다. 딥러닝이 널리 사용되기 시작하면서 SRCNN[1]을 필두로 하여, EDSR[2], RCAN[3], SAN[4] 등 괄목할 성능 향상을 보인 딥러닝 기반 초해상화 기술들이 소개되었다. 특히 최근 가장 뛰어난 성능을 보이는 기술 중 하나인 SAN 은 RCAN 과 기본 네트워크 구조는 비슷하지만 2 차 어텐션과 구역 단위의 비지역적 어텐션 (Region-level Non-local attention)을 추가 활용하는데, 전체 이미지에서 진행하기에는 연산량이 과하게 요구되어 RCAN 과 SAN 의 모델을 결합한 형태의 네트워크를 초해상화 네트워크로 활용하였다.

손실된 픽셀들을 맥락적, 구조적으로 자연스럽게 채워 넣는 인페인팅은 초해상화보다 더욱 복잡한 영상처리 기법이다. 전통적인 방식으로는 구조적으로 자연스럽게 만드는 PatchMatch[5] 등이 있는데, 맥락적인 자연스러움은 기대할 수 없다는 한계가 있었다. 딥러닝이 도입되며 DeepFill[6], HiFill[7], Zoom-to-Inpaint[8] 등의 기술들이 연구되었고, 이들은 구조와 맥락을 모두 고려하는 우수한 성능의 인페인팅 결과를 보여주었으며, [7]의 경우는 초해상화를 함께 적용하는 대신 고해상도의 영상에서 인페인팅을 가능하게 했다. 해당 논문에서는 더 많은 필터 수를 통해 많은 특징을 포착하고 각 스케일 단계마다 맥락적 어텐션(Contextual attention)의 정보를 활용할 수 있도록 [7]과 [8]에 기반하여 인페인팅 네트워크를 구성하였다.

다양한 복합적인 문제들에 대해 합동 학습을 하는 연구들이 이미 존재하지만 인페인팅과 초해상화를 동시에 수행하는 연구는 그 필요성에도 불구하고 거의 진행되지 않았다. 두 가지 문제를 해결하는 특성이 매우 다르다는 제약으로 인해 다른 합동 처리 문제들 보다 상당히 어려운 문제에 속한다. 주변 픽셀 정보들을 참조하는 형태인 초해상화와 그와는 달리 전체 영상의 정보에서 맥락적, 구조적 특징을 모두 학습해야 하는 인페인팅을 동시에 수행하는 것은 네트워크가 학습할 내용이 너무 많아지므로 오히려 성능 저해를 가져올 것으로 보이기도 한다.

본 논문에서는 간단한 연속적 모델과 제안하는 손실 함수를 통해 앞서 예측한 단점들이 극복될 수 있음을 입증한다. 복합 문제 해결의 가장 간단한 방식인 순차적으로 단일 모델을 활용하는 경우와 합동 처리 모델의 성능을 비교하여, 서로의 특성을 함께 학습함으로써 오히려 더 향상된 결과를 만들어 낼 수 있음을 보인다. 또한 각 단계마다 손실함수를 구하고, 기본 L1

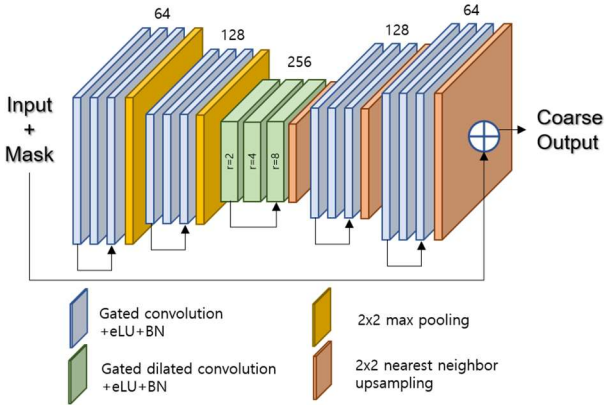

그림 2. 개략적 복원 네트워크(Coarse network) 구조

함수 외에 단계별로 적절한 손실 함수를 추가하는 다단계 결합 손실 함수를 제안한다. 본 논문에서 제안한 초해상화와 인페인팅 동시 수행의 해결 방안은 저화질에 스크래치나 얼룩을 포함하고 있는 오래된 사진을 복원하는 복잡한 문제 등 역시 한 번의 네트워크로 해결될 수 있다는 가능성을 제시한다.

2. 네트워크 구조

본 논문에서 사용되는 네트워크 구조는 그림 1 과 같은 3 단계의 네트워크로, 개략적 복원 네트워크, 디테일 복원 네트워크, 초해상화 네트워크로 구성되어 있다. 1, 2 단계는 인페인팅을 위한 네트워크, 3 단계는 초해상화를 위한 네트워크이다.

1 단계의 개략적 복원 네트워크는 원본 영상을 축소시킨 후 2 진 마스크 (binary mask)에서 값이 1 인 픽셀 부분을 지워낸 영상과 그 마스크를 함께 입력으로 받는다. 인페인팅 네트워크는 모두 게이트 컨벌루션 (Gated convolution) [9]으로 이루어져 있으며, 잔차 (Residual) 블록으로 구성된 인코더-디코더 형태이다.

2 번째 단계의 디테일 복원 네트워크는 앞의 결과를 입력 영상으로 하며, 역시 게이트 컨벌루션으로 구성되어 있는데 인코더의 끝 부분에서 [7]에서 보여준 방법과 동일한 맥락적

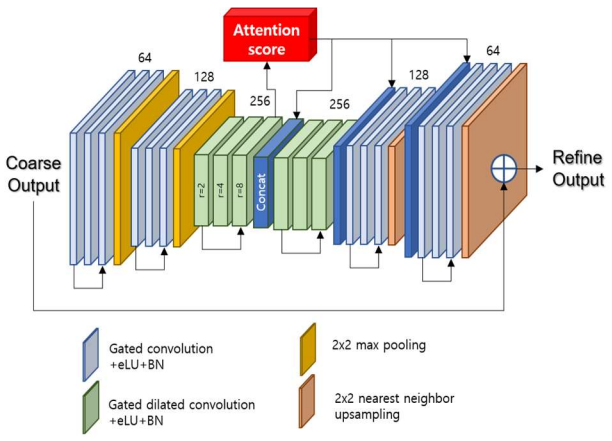

그림 3. 디테일 복원 네트워크(Refinement network) 구조

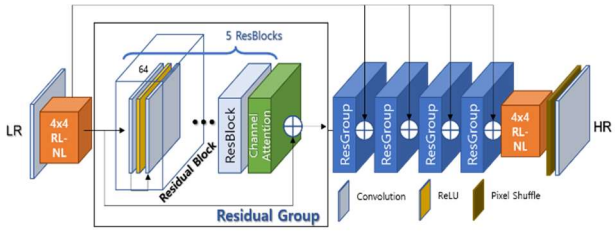

그림 4. 초해상화 네트워크 (SR network) 구조

어텐션을 사용하는데, 채워야 할 영역의 패치와 알고 있는 영역의 패치 사이에서 유사도를 구해 어텐션 정보를 디코더의 블록마다 결합하게 된다.

3 번째 단계의 초해상화 네트워크은 앞의 인페인팅 네트워크들과 달리 일반 컨벌루션을 활용하며 RCAN 의 기본 구조인 Residual Group 으로 구성되어 있다. 본 논문에서는 각 그룹은 5 개의 Residual 블록으로 구성되고, 전체 네트워크는 5 개의 그룹으로 이뤄지도록 했다. 또한 SAN 의 아이디어 중 하나인 구역별 비지역적 어텐션을 추가한 형태를 사용하였다.

3. 다단계 결합 손실 함수

그림 1 에서 볼 수 있듯 각 단계 별로 손실 함수를 따로 구하는데, 이 때 각 단계의 결과를 이미지 도메인으로 복원하면 스케일 별 정답 영상과 비교할 수 있게 되므로 이를 이용해 단계별 손실함수를 구한다. 이때 단계별로 손실함수를 다르게 하는데, 1 단계와 2 단계에서는 L1 손실함수와 VGG 손실함수[10]를 활용한다. VGG 손실함수는 영상에서 단순 픽셀보다 의미적인 부분을 보게 하는 함수로 인페인팅 네트워크의 학습에 큰 도움을 줄 수 있다. 3 단계는 초해상화 네트워크만 고려한다면 L1 손실함수로도 충분하지만, GAN 을 추가 적용함으로써 영상 전체의 디테일을 보다 충실히 학습하여 초해상화의 결과가 향상되도록 하고, 초해상화 네트워크를 통과하며 더욱 커진 마스크 영역에 GAN 이 적용되므로 인페인팅 네트워크가 더 큰 해상도에서 디테일을 만들어 낼 수 있게 하며 인페인팅 성능까지 개선시킨다. 이때 판별기는 PatchGAN[11]을 사용하며, GAN 손실 함수로는 WGAN 손실 함수[12]를 활용한다. 전체 손실 함수는 따라서 다음과 같다.

$$L_{total} = L1(out_{coarse}, target_{lr}) + \lambda_{vgg} * VGG(out_{coarse}, target_{lr}) + L1(out_{refine}, target_{lr}) + \lambda_{vgg} * VGG(out_{refine}, target_{lr}) + L1(out_{final}, target_{hr}) + \lambda_{adv} * Adv(out_{refine}, target_{hr})$$

이때 λ_vgg=0.01, λ_adv= 0.01 or 0.001 로 실험적으로 설정한다.

4. 학습 결과

4.1. 실험 환경 설정

모든 실험은 아담 최적화 기법을 사용하였고, 초기 학습률은 1e-4 로 설정하였으며, 모든 웨이트는 Xavier 초기화 방식을 사용하였다. 데이터는 모든 데이터가 256x256 크기이며, 365 장면의 카테고리로 180 만 장의 학습 이미지를 포함하는 Places2-256[13]을 사용하였으며, 정답 영상을 원본 이미지, 이를 Bicubic interpolation 으로 2 배 축소한 128x128 영상을 작은 스케일에서의 정답 영상으로, 이에 랜덤 마스크를 더한 이미지를 입력 영상으로 하여 배치 사이즈는 8 이다.

각 네트워크는 전체 학습을 거치기 전 네트워크 별로 개략적 복원 네트워크 5 만, 디테일 복원 네트워크 5 만, 초해상화 네트워크 10 만번의 반복만큼 미리 학습을 시켰으며, GAN 학습을 적용시키기 전 충분히 수렴할 만큼 미리 학습을 시켰고, GAN 학습을 적용시킨 후 각각 학습을 모델별로 40 만~60 만 번 사이의 반복만큼 수렴할 때 까지 추가로 학습시켰다.

총 4 개의 비교 실험을 진행하였는데, 순차적 단일 모델 실험과 합동 처리 모델 실험을 각각 적대적 손실의 계수가 0.01 과 0.001 인 경우에 대해 진행하고 결과를 비교하였다. 테스트 결과 비교는 동일한 Places2-256 의 테스트 데이터 셋에서 앞 1000 개의 데이터에 대한 결과의 평균으로 진행하였다.

4.2. 실험 결과

	PSNR ↑	SSIM ↑	FID ↓
Sequential (λ_{adv}=0.01)	28.4940	0.8909	40.2789
Sequential (λ_{adv}=0.001)	28.4624	0.8897	40.3782
Joint (λ_{adv}=0.01)	27.2992	0.8609	38.1810
Joint (λ_{adv}=0.001)	28.7127	0.8953	38.1805

표 1. 정량적 성능 결과 비교

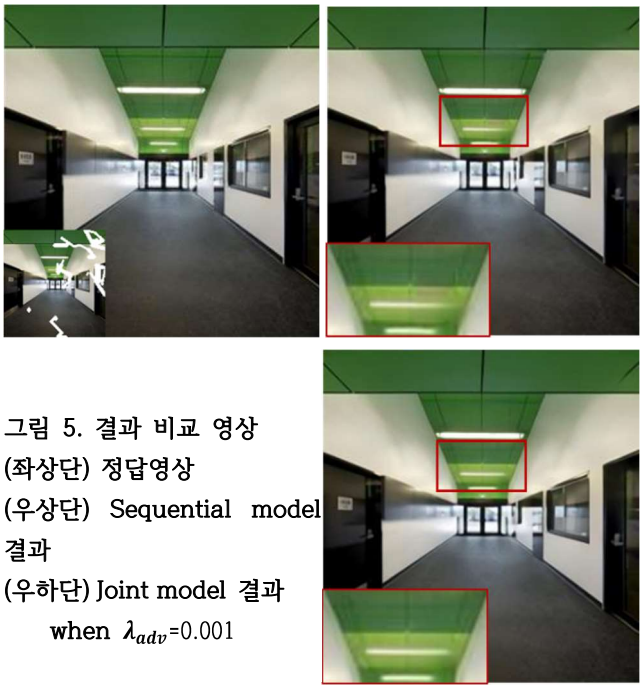

그림 5. 결과 비교 영상
(좌상단) 정답영상
(우상단) Sequential model 결과
(우하단) Joint model 결과
when λ_{adv}=0.001

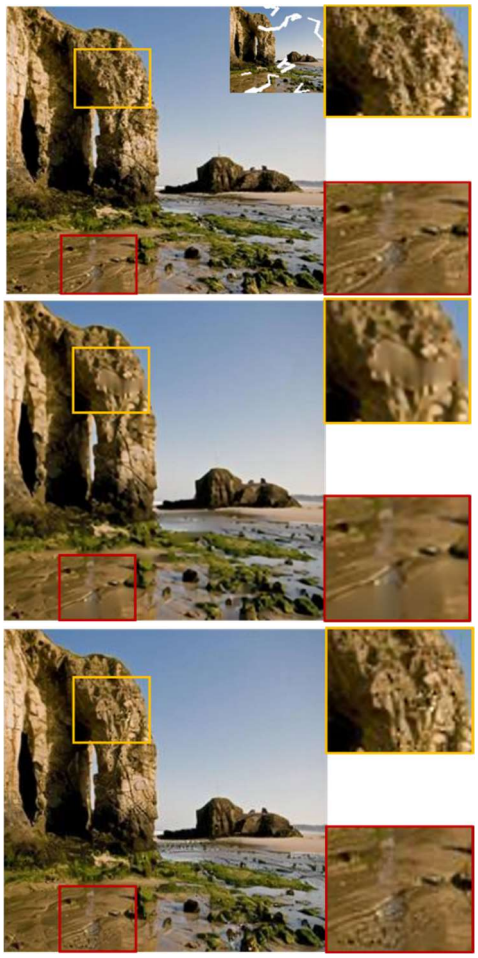

그림 6. 결과 비교 영상 (a) 정답 영상 (b) Sequential model 결과 (c) Joint model 결과 when $\lambda_{adv} = 0.01$

표 1 을 보면 adversarial 손실 함수의 계수가 0.001 일 때 합동 처리 모델이 PSNR, SSIM, FID 수치에서 모두 가장 우수한 것을 확인할 수 있다. 특히 인지적 화질을 나타내는 FID 수치의 경우 합동 처리 모델의 결과가 최대 2.19 가량 개선되었다. 이는 인페인팅과 초해상화 성능들이 모두 단일 모델보다 합동 모델에서 더욱 우수함을 입증한다. 또한 그림 6 을 확인하면 형광등 부분이 순차적 단일 모델의 경우보다 영상의 구조적 특성이 보다 선명함을 확인할 수 있다. Adversarial 손실 함수의 계수가 0.01 인 경우에서는 PSNR, SSIM 수치는 높지 않지만 FID 값은 높게 도출되었는데, 실제로 결과 영상을 확인 했을 때 디테일을 만들어내는 능력이 가장 뛰어났다. GAN 기반 모델들의 경우 자연스럽고 선명한 디테일을 생성하지만 이 경우 정답 영상과 정확히 일치하기는 어려울 수 있기에 L1 손실함수 기반 평가지표인 PSNR 과 SSIM 에 대해서는 낮은 값이 나왔을 것이라 예측한다.

5. 결론

순차적으로 단일 모델을 활용하는 방식에서는 인페인팅이 잘못될 경우 그 에러가 초해상화 네트워크를 거치며 더욱 커지게 되는데 비해, 본 논문의 제안 방법인 합동 처리 모델에서는 그러한 부작용을 막을 수 있었다. 또한 학습에 있어 상당히 다른 특성을 갖는 기법들인 인페인팅과 초해상화를 함께 학습하더라도 두 종류의 기법에서 모두 더욱 우수한 결과를 보일 수 있다는 것을 보였다. 추후에는 화질 저하와 스크래치같은 다양한 손상이 포함되어 있는 오래된 사진을 복원하는 등 다양한 실제 문제에서 활용할 수 있을 것으로 기대된다.

Acknowledgement

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 연구개발지원사업으로 수행되었음(과제번호: R2020040045)

참고 문헌

[1] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution”, In ECCV, 2014.

[2] B. Lim, S. Son, H. Kim, S. Nah and K.M. Lee, “Enhanced Deep Residual Networks for Single Image Super-Resolution”, In CVPRW, 2017.

[3] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong and Y. Fu, “Image Super-Resolution Using Very Deep Residual Channel Attention Networks”, In ECCV, 2018.

[4] T. Dai, J. Cai, Y. Zhang, S. Xia and L. Zhang, “Second-Order Attention Network for Single Image Super-Resolution”, In CVPR, 2019.

[5] C. Barnes, E. Shechtman, A. Finkelstei and, D. B. Goldman, "PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing.", In ACM Transactions on Graphics (Proc. SIGGRAPH) 28(3), August 2009.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. S. Huang, “Generative Image Inpainting with Contextual Attention”, In CVPR, 2018.

[7] Z. Yi, Q. Tang, S. Azizi, D. Jang and Z. Xu, “Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting”, In CVPR, 2020.

[8] S. Kim, K. Aberman, N. Kanazawa, R. Garg, N. Wadhwa, H. Chang, N. Karnad, M. Kim and O. Liba, “Zoom-to-Inpaint: Image Inpainting with High-Frequency Details”, In arXiv, 2021.

[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. Huang, “Free-Form Image Inpainting with Gated Convolution”, In ICCV, 2019.

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., “Photo-realistic single image super-resolution using a generative adversarial network”, In CVPR, 2017.

[11] P. Isola, J. Zhu, T. Zhou and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks”, In CVPR, 2017.

[12] M. Arjovsky, S. Chintala and L. Bottou, “Wasserstein GAN”, In PMLR, 2017.

[13] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million Image Database for Scene Recognition”, In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.