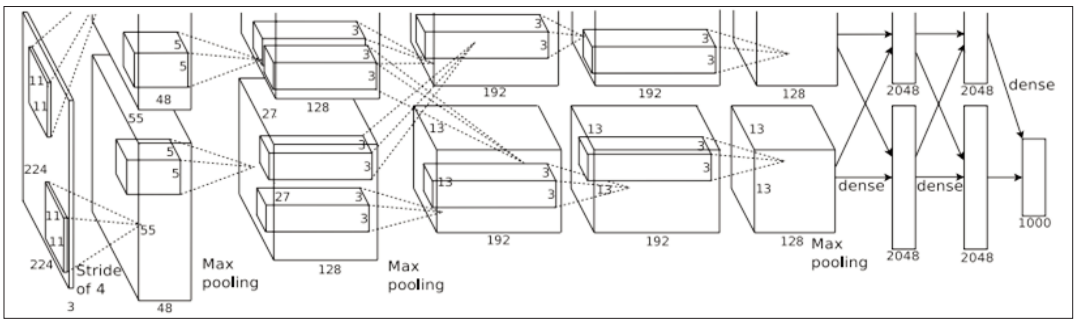
Deep Learning for Low-Level Computer Vision

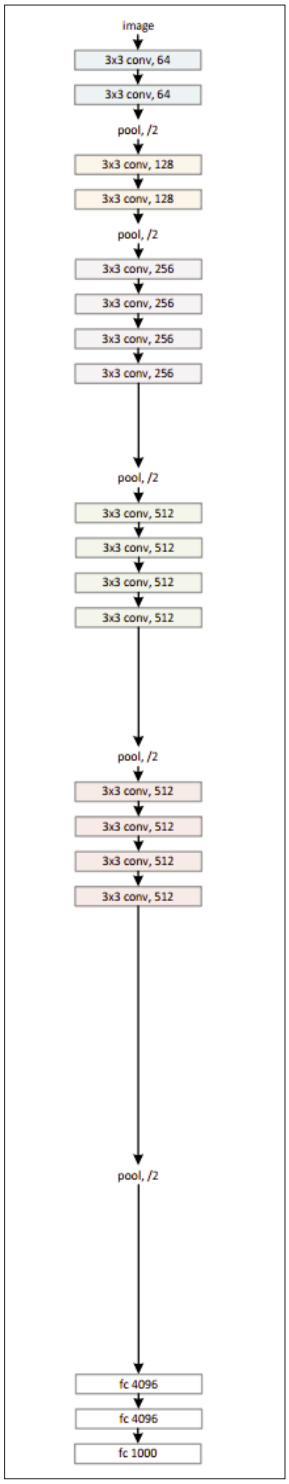
□ 오승욱, 조영현, 김선주 / 연세대학교

Ⅰ. Introduction

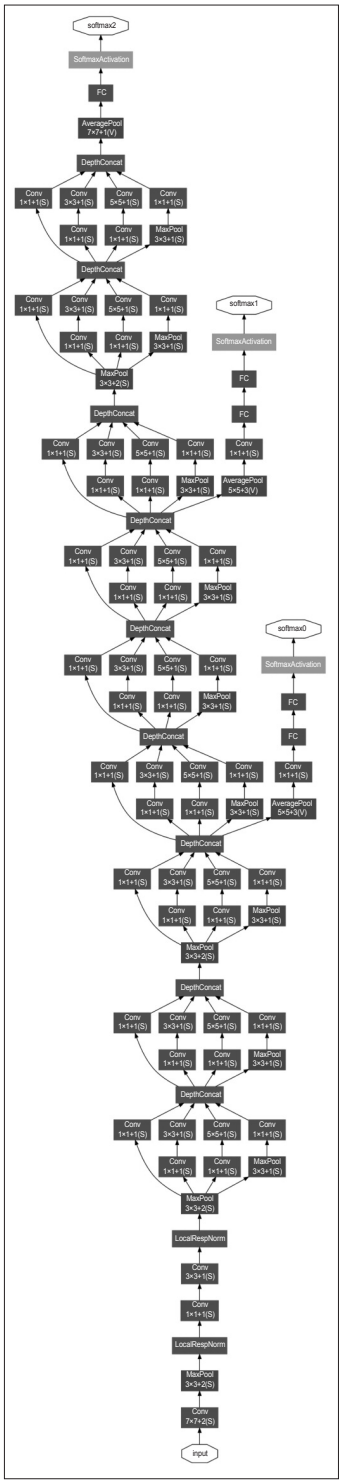
딥러닝, 특히 컴퓨터비전 분야에서 가장 활발히 이용되고 있는 이 기술은 의미 있는 성과를 낸 지 4-5년 정도 된 상대적으로 최신 기술이다. 컴퓨터 비전 분야에서 딥러닝을 촉발시킨 연구는 AlexNet[1]으로, 이는 영상인식(Image classification)에 이용된 Convolutional Neural Network(CNN)이다. AlexNet은 영상인식 대회인 ImageNet Challenge[2] 2012년 classification 부분에서 딥러닝을 이용하지 않은 나머지 팀들을 압도적으로 따돌리며 우승했다. 물론 이전에도 인공신경망(Artificial Neural Network)이라는 기술이 존재했다. 하지만 여러 가지 제약으로 인해 한동안 잠잠했던 이 기술이, 병렬 컴퓨팅 성능의 향상과 폭발적으로 늘어난 데이터로 딥러닝이라는 이름으로 부활했다.



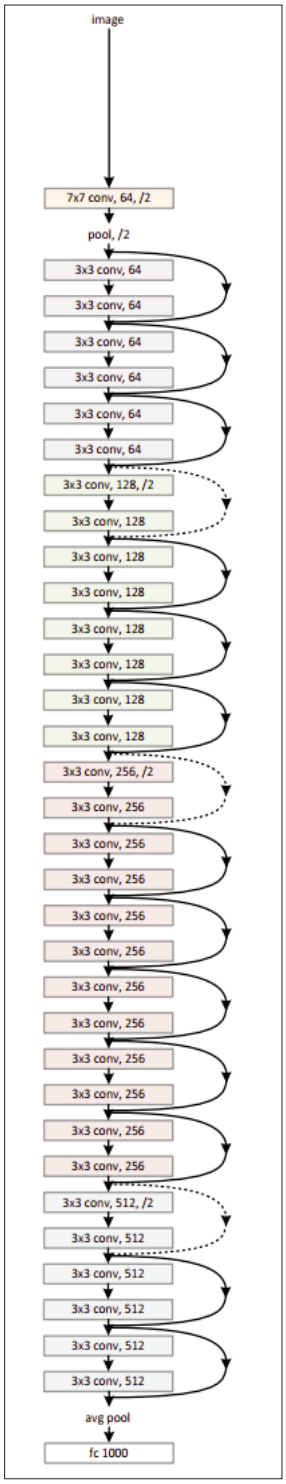
〈그림 1〉 AlexNet[1], 8 layers



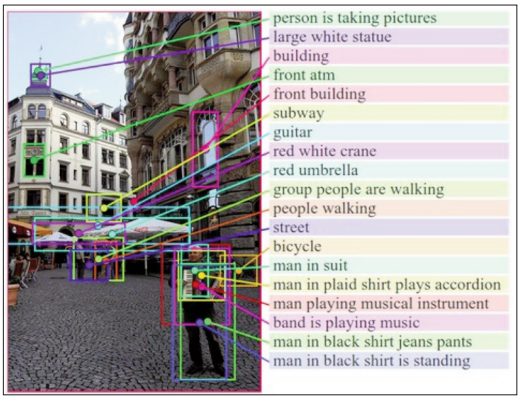
〈그림 2〉 VGG[3], 19 layers



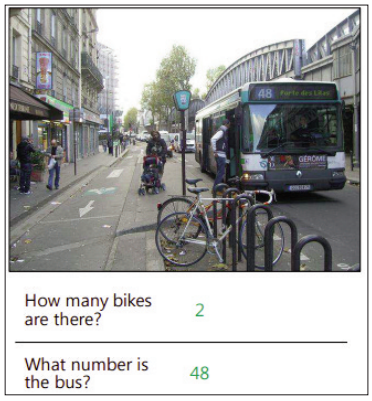
〈그림 3〉 GoogLeNet[4], 22 layers



〈그림 4〉 ResNet[5], 34 layers



〈그림 5〉 이미지 captioning[6] 예시



〈그림 6〉 VQA[7] 예시

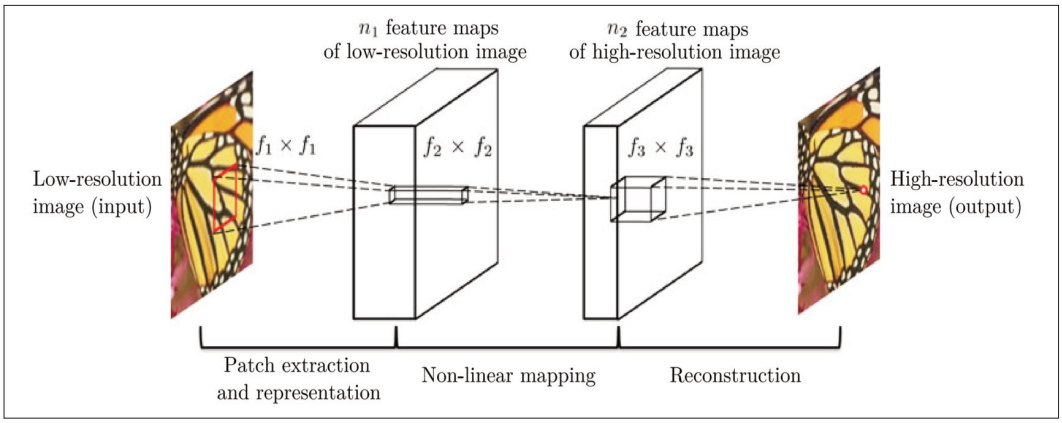
이를 계기로 수많은 연구자들이 다양한 컴퓨터비전 문제들을 해결하기 위해 딥러닝을 이용한 연구에 뛰어들었다. 기존의 방식으로는 풀기 어려웠던 문제들을 data-driven 방식의 딥러닝을 이용하여 풀기 시작했다. 영상인식의 경우 AlexNet 이후로 성능 향상을 위해 더 깊은(Deep) 구조를 갖는 CNN을 만들어 이용했다. 대표적으로 VGG[3], GoogLeNet[4], ResNet[5] 등이 있다. 현재 딥러닝을 이용한 영상인식의 성능은 인간과 비슷한 수준의 능력까지 도달했다. 최근에는 딥러닝으로 단순 영상인식 뿐만 아니라 Image captioning[6], Visual Question Answering(VQA)[7] 등과 같은 영상이해 측면의 high-level 컴퓨터비전 문제를 해결하고 있다.

또한 Image super-resolution[8], Image deblurring[9] 등과 같이 영상을 입력으로 받아 처리된 영상을 출력하는 low-level 컴퓨터비전 문제에도 딥러닝은 좋은 결과를 보여주고 있다. 본 기고에서는 특히 low-level 컴퓨터비전 분야에 적용된 딥러닝 사례들을 살펴보고자 한다. 현재까지 나온 많은 논문을 보면 딥러닝이 high-level 컴퓨터비전 분야뿐만 아니라, 다른 컴퓨터비전 분야에도 충분히 활용될 수 있다는 것을 보여주고 있다.

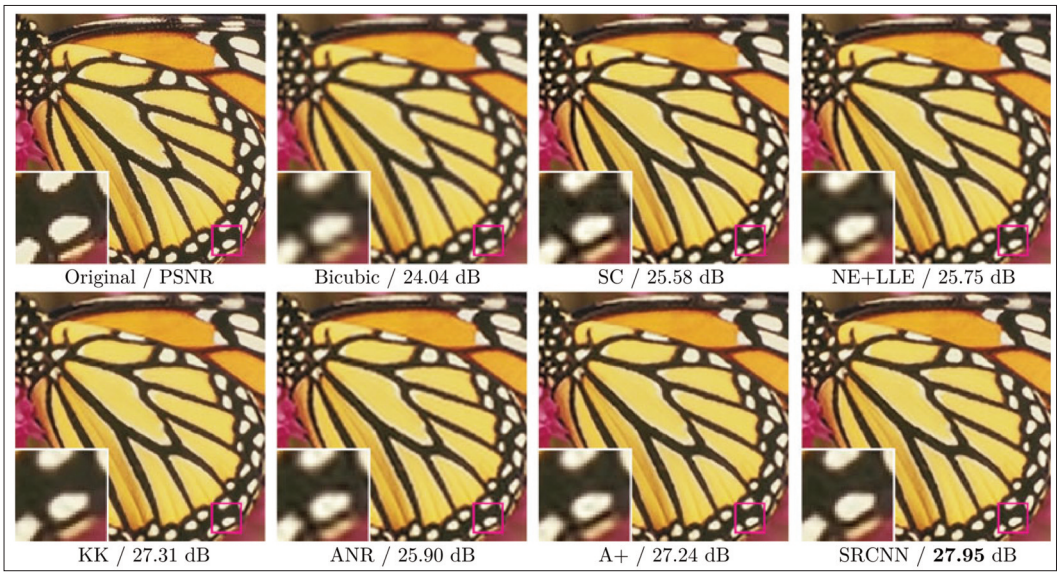
II. Deep Learning for Image Super-Resolution

영상 초고해상도 복원 기술은 영상처리 및 컴퓨터비전 분야에서 가장 활발히 연구되는 분야 중 하나이다. 2000년대 들어 Sparse coding과 같은 data-driven 방식으로 영상 초고해상도 복원 성능을 향상시키려는 연구가 많았지만, 최근에는 딥러닝을 이용하면서 성능이 월등히 향상됐다. 딥러닝을 이용한 영상 초고해상도 연구를 촉발시킨 것은 SRCNN[8]이다. SRCNN은 세 개의 convolution layer로 이루어진 간단한 구조이지만, 성능은 딥러닝을 이용하지 않은 다른 알고리즘보다 뛰어나다.

SRCNN의 첫 번째 convolution layer는 저해상도 입력 영상에서 패치별 특징을 추출한다. 두 번째 convolution layer는 1×1 필터를 이용하여 각 패치별 특징벡터를 고해상도 패치에 대한 특징벡터



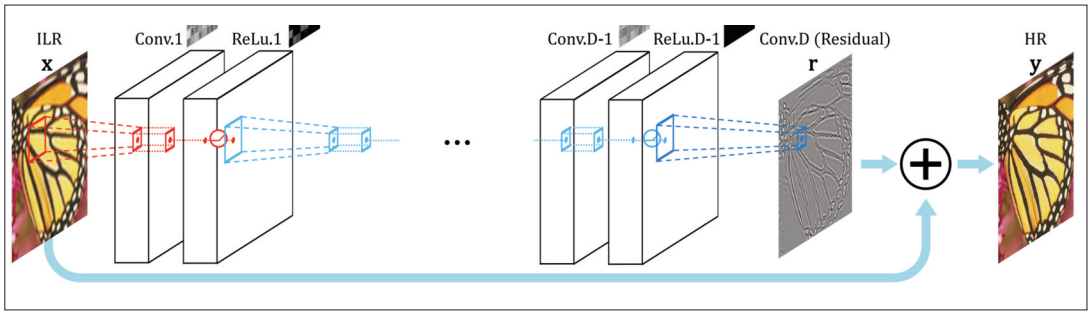
〈그림 7〉 SRCNN[8] 구조



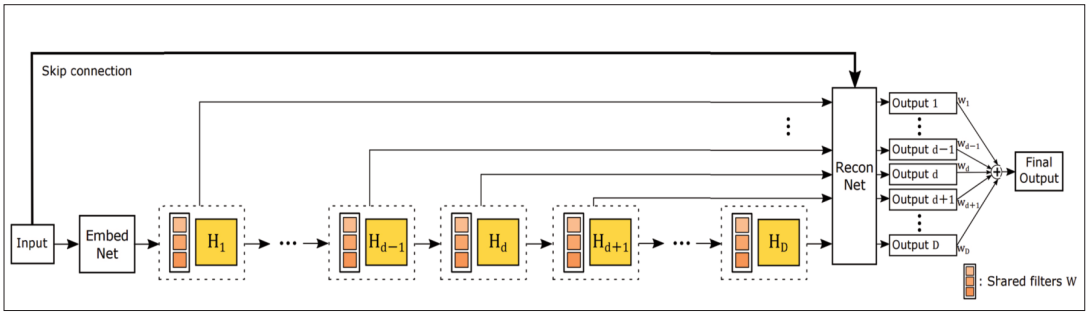
〈그림 8〉 SRCNN 결과

로 매핑한다. 그리고 마지막 convolution layer를 통해 고해상도 영상을 복원하는데, 이 때 필터 크기를 1×1보다 크게 함으로써 이웃한 패치 사이에 자연스러운 연결을 고려한 결과를 얻을 수 있다.

영상 초고해상도 복원의 경우도 SRCNN 이후로 성능 향상을 위해 더 깊은 구조를 갖는 CNN이 등장했다. 대표적으로 VDSR[10]과 DRCN[11]이 있다. VDSR은 20개의 convolution layer를 쌓아 성능을 향상시켰는데, 깊은 구조를 잘 학습시키기 위해 skip-connection을 이용했다. DRCN도 이와 유사하게 20개의 convolution layer를 갖지만 네트워크의 파라미터를 재사용함으로써 구조의 효율을 높였으며, recursive-supervision과 skip-connection을 이용하여 영상 초고해상도 복원 성능을 높였다.



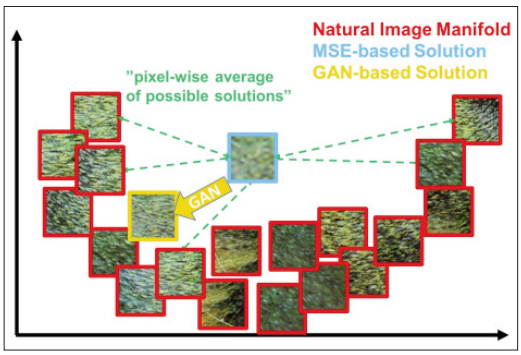
〈그림 9〉 VDSR[10] 구조



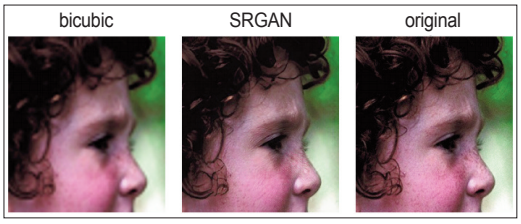
〈그림 10〉 DRCN[11] 구조

하지만 SRCNN, VDSR, DRCN을 포함한 대부분의 딥러닝을 이용한 영상 초고해상도 복원 기술의 단점이 있다. 영상 초고해상도 성능 평가의 기준이 Peak Signal-to-Noise Ratio(PSNR)로 이루어지기 때문에, 딥러닝 네트워크의 최적화에 있어서 대부분의 경우 cost function을 Mean Squared Error(MSE)를 이용한다. MSE를 이용하면 전체적인 에러는 줄일 수 있지만, 동시에 결과 영상이 부드럽게 되어 버린다. 즉 PSNR은 높아졌지만, 인간이 직접 보고 느끼는 시각 인지적인 측면에서는 좋은 결과가 아닐 수 있다.

이를 해결하기 위해 Generative Adversarial Network(GAN)를 이용한 영상 초고해상도 복원 연구가 진행되고 있다. 대표적으로 SRGAN[12]이 있으며, 원리는 저해상도의 입력 영상을 학습된 natural image manifold에 최대한 근접하도록 만들어 주어 고해상도의 출력 영상이 natural image와 유사하게 보이도록 만드는 것이다.



〈그림 11〉 SRGAN[12] 원리



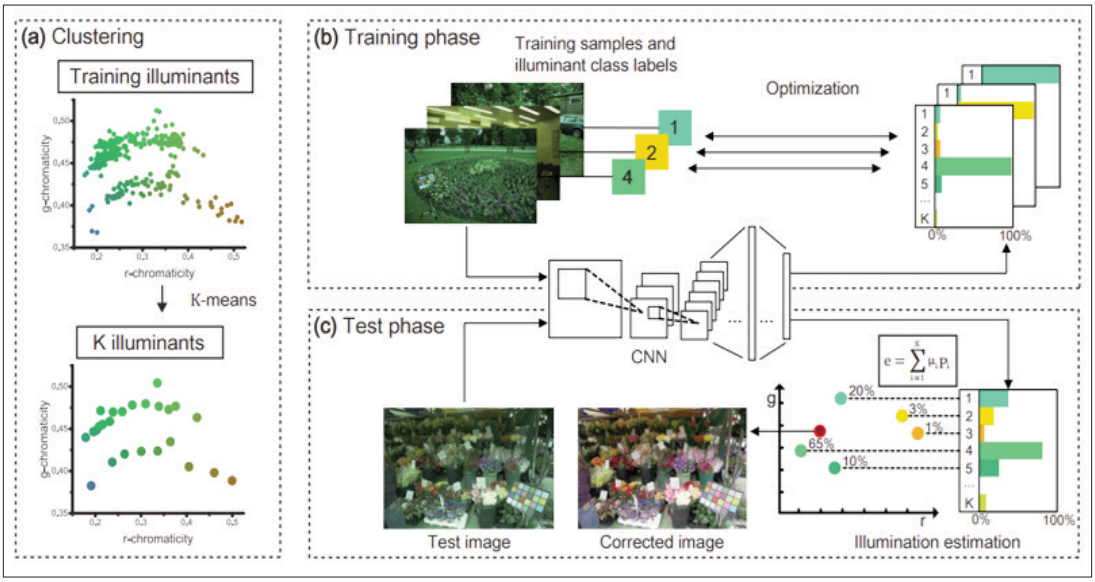
〈그림 12〉 SRGAN 결과

이전에도 PSNR과 Human Visual System(HVS) 간의 상관관계가 높지 않다는 연구들이 있었다[13]. GAN을 이용한 영상 초고해상도 복원 결과를 통해서도 단순히 PSNR이 높다고해서 인간의 눈에 좋게 보이는 결과가 아니라는 것이 또 한 번 입증된 셈이다. 이처럼 앞으로의 영상 초고해상도 복원 문제는 단순히 PSNR을 높이는 것이 아니라, 인간의 시각 인지적 욕구를 만족시키는 방향으로 발전할 것이다.

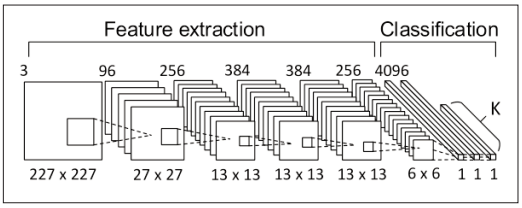
Ⅲ. Deep Learning for Computational Color Constancy

우리의 눈으로 들어오는 빛은 물체 본연의 색과 조명의 색이 섞인 상태이다. 하지만 인간의 시각시스템에는 조명의 색과 물체의 색을 분리하는 능력이 있어 다양한 조명 환경에서도 물체의 색을 정확하게 인식할 수 있다. 이러한 능력을 색상 항상성(Color constancy)이라고 한다.

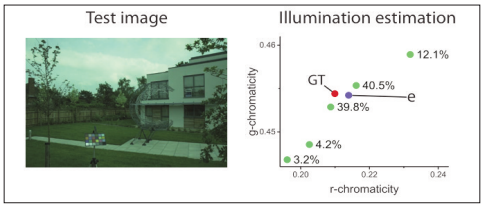
하지만 문제는 카메라는 이러한 기능이 없다는 것이다. Computational Color Constancy란 다양한 조명 조건에서 카메라로 촬영된 영상에서 조명의 색과 물체의 색을 분리하여 정확한 물체의 색을 구해내는 것을 목표로 하는 문제이다. 즉, 실제로 흰색 물체가 사진상에서도 흰색으로 보이도록 하는 것을 말한다. 카메라에서는 이러한 기능을 화이트 밸런싱(White Balancing)이라고 한다. 사진 속 물



〈그림 13〉 Deep learning for Computational Color Constancy[14] 알고리즘 개념도



〈그림 14〉 학습에 사용한 네트워크 구조



〈그림 15〉 조명 분류 정보를 이용한 조명 예측 과정

체의 정확한 색을 구해내기 위해서는 조명에 대한 정보를 사진으로부터 유추해 내야 한다. 기존에는 단순히 영상의 평균 혹은 최댓값이 조명의 색과 유사하다는 가정인 Gray-world, White patch 등의 방법을 사용하여 조명의 색을 예측하였지만, 최근에는 이 문제 역시 Data-driven 접근으로 딥러닝을 많이 활용하는 추세이다.

필자의 연구실 역시 딥러닝 방법을 활용해 기존의 방법보다 더욱 정확하게 조명의 색을 예측할 수 있는 방법을 개발하여 발표했다[14]. 이 연구에서 새롭게 제안한 아이디어는 다음과 같다. 조명의 색은 연속적인 공간에 존재하기 때문에 기존 방법에서는 회기(regression) 방법으로 조명의 색을 예측하는 경우가 대부분이다. 하지만, 이 연구에서는 무수히 많이 존재하는 비슷한 조명들을 그룹화 시켜서 분류(classification)를 하는 딥네트워크를 학습했을 경우 기존보다 훨씬 더 정확하게 조명을 예측할 수 있다는 것을 보여준다. 이러한 방법을 사용하면 조명을 예측하는 딥네트워크의 정확도를 올릴 수 있다. 그 다음으로 딥네트워크를 통해 예측한 조명의 분류정보를 이용하여 다시 연속적인 원래의 조명색 공간으로 매핑하는 방법을 통해서

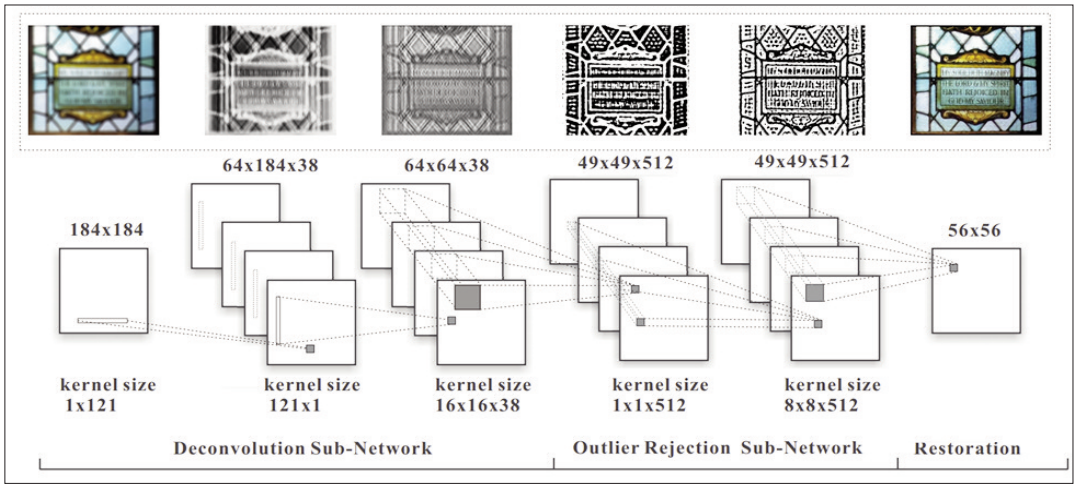
〈그림 16〉 기존의 방법과 제안한 방법을 통한 조명 예측 및 예측된 조명을 이용한 화이트밸런싱 결과. 우측하단의 값은 예측된 조명의 에러값

정확한 조명을 예측할 수 있는 방법을 고안했다.

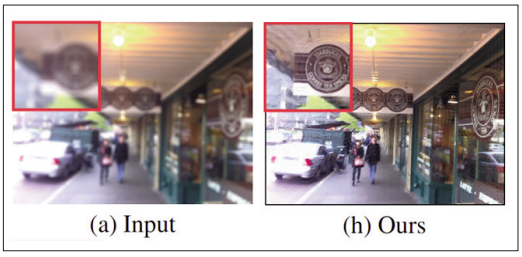
학습에 사용된 네트워크는 기존의 영상 인식에 사용된 네트워크 구조에 마지막 레이어의 출력을 조명 그룹의 개수로 수정하여 사용하였다. 딥러닝의 특성상 많은 학습데이터를 필요로 하지만, 조명 예측을 학습할 수 있는 데이터셋은 한정돼 있다는 문제가 있다. 이를 해결하기 위해 전이학습 (transfer learning) 기법을 사용하였다. 많은 학습 데이터가 존재하는 영상 인식을 위해 먼저 네트워크를 학습한 후, 학습된 필터를 이 문제를 위해 Fine-tuning하는 방식을 통해 안정적으로 적은 데이터를 통해 학습이 가능했다. 추가적으로, Data augmentation을 통해서 임의로 회전, 이동, 뒤집는 변환을 통해서 적은 데이터에 overfitting되는 것을 방지했다.

Ⅳ. Deep Learning for Image Deblurring

이미지 디블러링이란 카메라 혹은 장면의 움직임에 의해 흐려진 영상을 복원하는 작업을 일컫는다. 보통 컨벌루전 모델을 사용해서 이미지 디블러링을 모델링하며 블러 커널의 역연산을 예측하여 선명한 원본 영상을 복원하는 것을 목표로 한다. 최근 영상 디블러링 분야에서도 딥러닝이 활발히 활용되고 있는데, 대표적인 알고리즘으로 DCNN[9]가 있다. 이 연구에서는 특정한 블러 커널에 의해서 흐려진 영상을 복원하는 CNN을 학습하는 것을 목표로 하는데, 이를 위해 특별한 CNN 구조를 설계하였다. 선명한 영상을 얻기 위해서는 컨벌루젼 연산의 역연산을 학습하여야 하는데, 문제는 역연산을 위한 필터의 크기가 매우 크다는 것이다. 이 논문에서는 역커널(inverse kernel)의 가로 방향과 세로 방향의 연산을 분리할 수 있다는 kernel separability 특성을 활용한 네트워크 구조를 제안하였다. 제안한 네트워크는 가로로 긴 1×121 커널과 세로로 긴 121×1 커널을 사용하여, 121×121의 역커널을 근사할 수 있는 구조를 가지고 있다. 하지만, 이 네트워크는 특정 블러 커널을 위해 학습되며 그 커널로 흐려진 이미지만 복원할 수 있다는 한계점을 가지고 있다.



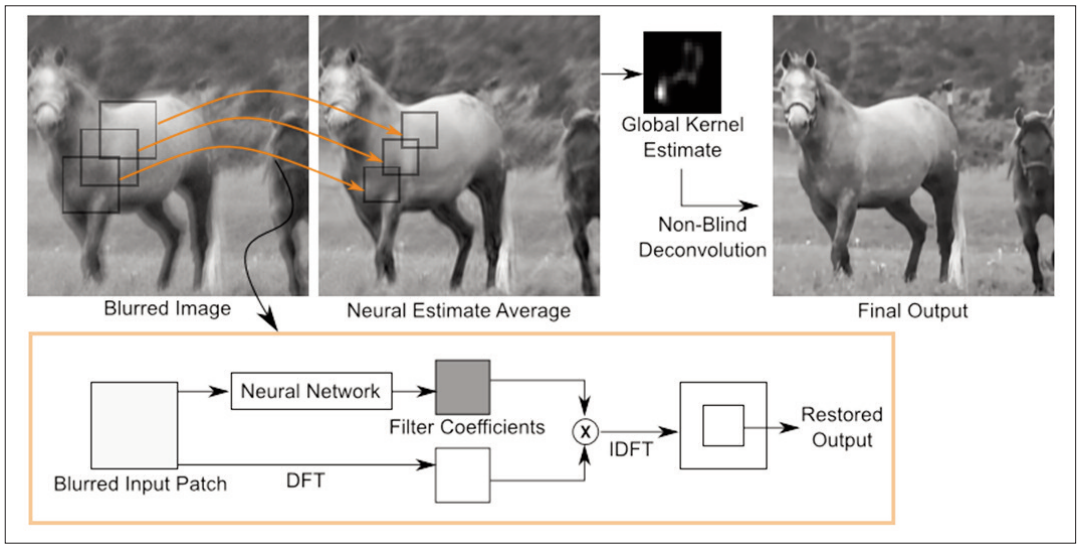
〈그림 17〉 DCNN[9]의 네트워크 구조



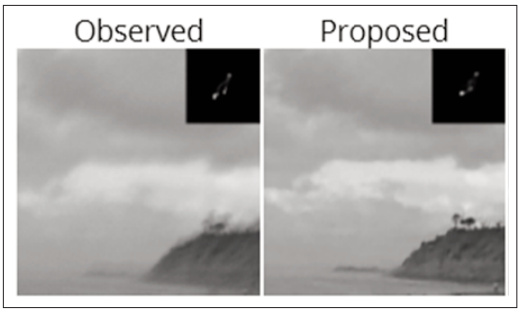
〈그림 18〉 DCNN[9]의 결과

이러한 DCNN의 한계를 극복하기 위해 모든 블러 커널에 사용가능하며, 블러 커널에 대한 정보 없이도 동작하는 영상 디블러링 알고리즘인 NDeblur[15]가 최근에 제안되었다. 이를 가능케하는 가장 큰 원동력은 영상의 주파수 영역에서 학습된 딥네트워크이다. 이 논문에서는 영상의 주파수 영역에서 역커널의 계수를 예측하는 네트워크를 학습하여 흐려진 영상을 주파수 영역에서 복원한 후 다시 원래의 공간으로 매핑하는 방법을 제안했다. 작은 패치 단위로 네트워크를 학습하였으며, 테스트 시에는 패치의 평균이미지를 이용하여 최종 블러커널을 예측한 후 기존의 Non-blind deconvolution 방법을 사용하여 최종적으로 선명한 영상을 얻어냈다.

앞서 소개한 알고리즘들은 모두 컨벌루젼 모델을 사용하여 영상 블러 현상을 모델링하였다. 하지만, 실제 영상은 훨씬 더 복잡한 방식으로 흐려지게 되고 컨벌루젼 모델로는 실제 영상을 모두 설명할 수 없다. 즉, 앞서 소개한 알고리즘들은 실제 영상에서 제대로 작동한다는 보장이 없다는 것이다. 컨벌루젼 모델로 설명하기 힘든 현상으로는, 빠르게 지나가는 자동차가 흐려지는 것(Dynamic motion)과 카메라 회전 등이 있다. 이러한 문제를 해결하기 위해 최근 블러 모델 없이 동적 모션에 의해 흐려진 영상과 선명한 영상의 쌍을 취득하고 이를 학습하여 모델 제한없이 선명한 영상을 복원하는 방법에 대한 연구가 진행되고 있다[16]. 이 데이터를 취득하기 위해 frame rate가 높은 액션 캠을 사용하여 동적



〈그림 19〉 NDeblur[15] 알고리즘의 개념도

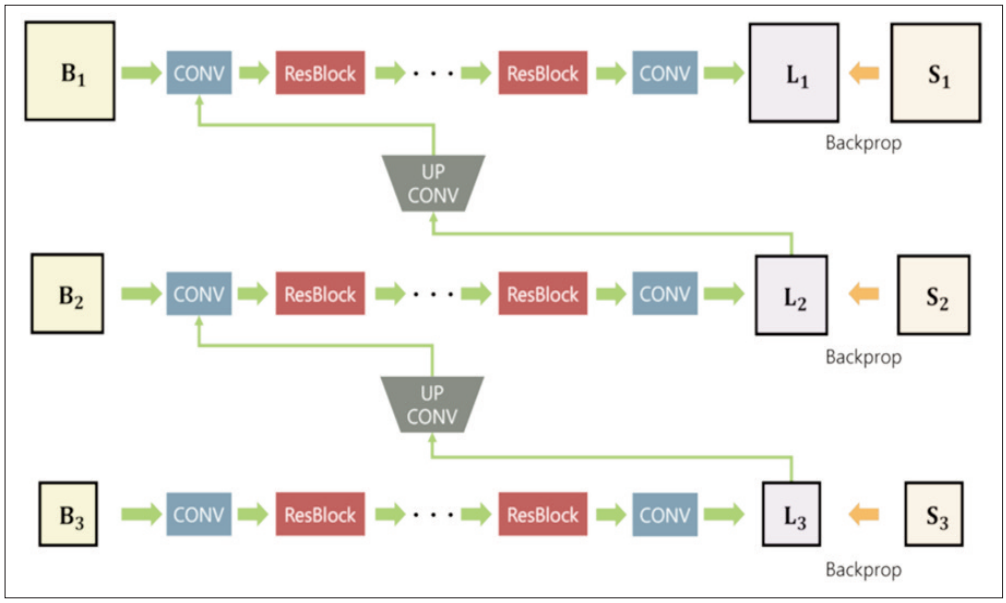


〈그림 20〉 NDeblur의 결과

인 장면을 촬영하고, 이를 시간축으로 적분하여 흐려진 영상을 얻어냈다. 이렇게 함으로써, 영상에서 특정한 모델을 가정하지 않고 디블러 알고리즘을 학습할 수 있는 데이터베이스를 구축하였다. 이 연구에서 사용된 네트워크 구조는 다중 스케일의 입력을 받아서 각각의 스케일에서의 결과를 합쳐서 최종적으로 선명한 영상을 복원할 수 있는 네트워크를 제안하였고, Adversarial 학습을 통해서 결과 영상이 실제 선명한 영상과 유사하도록 만들었다.

V. Image Colorization

Image Colorization 문제는 흑백 영상을 입력으로 받아, 컬러 영상을 복원해내는 문제를 말한다. 최근 딥러닝을 이용하여 이 문제를 해결한 연구가 발표되었다[17]. 이 연구에서는 영상의 고수준의 의미를 이용하여 이를 컬러 복원에 이용할 수 있도록 네트워크를 설계하였다. 구체적으로는 영상 인식에 사용되는 global feature와 mid-level feature를 통합하여 영상의 컬러 정보를 복원하는데 사용하였다. 이미지로부터 global feature를 뽑아낼 수 있도록 네트워크를 학습하기 위해 영상 분류와 컬러복원을 동시에 학습하는 멀티태스크 학습 방법을 적



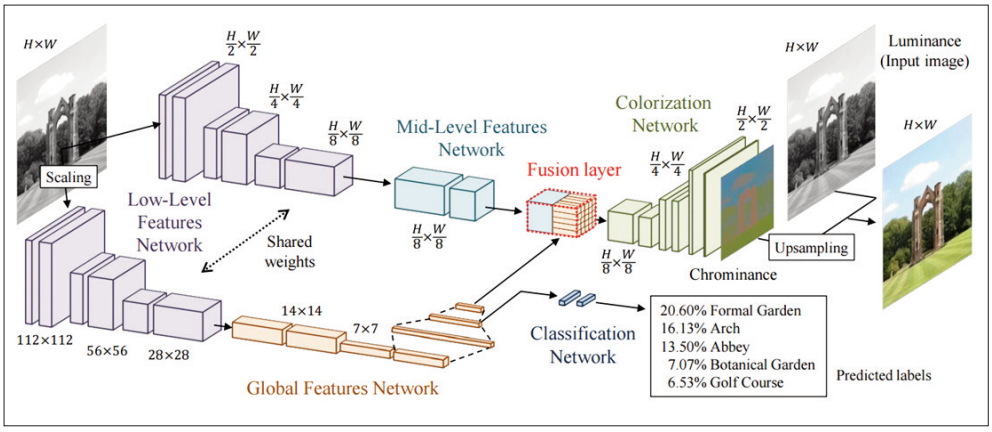
〈그림 21〉 멀티 스케일 네트워크 구조[16]

〈그림 22〉 Adversarial 학습의 효과

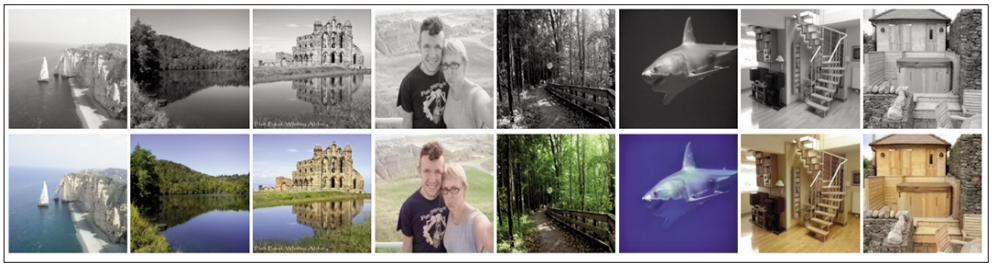
용하였다. 결과적으로는 흑백 영상의 고수준의 정보를 바탕으로 더욱 정확하게 컬러값을 예측할 수 있게 되었다.

VI. Discussion

딥러닝은 컴퓨터비전 분야에서 어떤 문제를 해결하는 방식을 본질적으로 바꾸어 놓았고 그 활용 분야 역시 점점 넓어지고 있는 추세이다. 또한 빠른 속도로 새로운 딥러닝 네트워크 구조 및 학습 패러다임이 개발되고 있다. 대표적인 예로는 최근에 제안된, 이미지를 출력으로 하는 네트워크의 성능을 급격하게 상승시킨 Adversarial 학습 방법(실제 영상과 생성된 영상을 구분하지 못하도록 학습)이 있다. 이와 같은 추세로 컴퓨터비전 분야의 더욱 어렵고 복잡한 문제들도 딥러닝으로 해결할 수 있을 것으로 예상된다.



〈그림 23〉 Image Colorization 알고리즘[17]의 네트워크 구조. 컬러 정보와 영상 분류를 동시에 학습



〈그림 24〉 Image Colorization 알고리즘[17]의 결과. 입력 흑백 영상(윗줄), 결과 컬러 영상(아랫줄)

참 고 문 헌

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[2] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[4] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." arXiv preprint arXiv:1512.03385 (2015).

[6] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[7] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[8] Dong, Chao, et al. "Image super-resolution using deep convolutional networks." IEEE transactions on pattern analysis and machine intelligence 38.2 (2016): 295-307.

[9] Xu, Li, et al. "Deep convolutional neural network for image deconvolution." Advances in Neural Information Processing Systems. 2014.

[10] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[11] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee. "Deeply-Recursive Convolutional Network for Image Super-Resolution." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[12] Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." arXiv preprint arXiv:1609.04802 (2016).

[13] Yang, Chih-Yuan, Chao Ma, and Ming-Hsuan Yang. "Single-image super-resolution: a benchmark." European Conference on Computer Vision. Springer International Publishing, 2014.

[14] Seoung Wug Oh and Seon Joo Kim. "Approaching the Computational Color Constancy as a Classification Problem through Deep Learning." Pattern Recognition, Volume 61, January 2017.

[15] Ayan Chakrabarti. "A Neural Approach to Blind Motion Deblurring" European Conference on Computer Vision. Springer International Publishing, 2016.

[16] Nah, Seungjun, Tae Hyun Kim, and Kyoung Mu Lee. "Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring." arXiv preprint arXiv:1612.02177 (2016).

필 자 소 개

오 승 욱

– 2014년 : 연세대학교 컴퓨터과학과 졸업(학사)
– 2014년 ~ 현재 : 연세대학교 컴퓨터과학과 석박통합 과정
– 주관심분야 : 컴퓨터 비전, 딥러닝

필 자 소 개

조 영 현

– 2015년 : 연세대학교 컴퓨터과학과 졸업(학사)
– 2015년 ~ 현재 : 연세대학교 컴퓨터과학과 석박통합 과정
– 주관심분야 : 컴퓨터 비전, 딥러닝

김 선 주

– 1997년 : 연세대학교 전자공학과 졸업(학사)
– 2001년 : 연세대학교 전기전자공학과 졸업(석사)
– 2008년 : 미국 Univ. of North Carolina at Chapel Hill 졸업 (박사)
– 2009년 ~ 2011년 : National University of Singapore (Research Fellow)
– 2012년 : 한국뉴욕주립대학교 (조교수)
– 2013년 ~ 현재 : 연세대학교 컴퓨터과학과 (조교수)
– 주관심분야 : 컴퓨터비전, Computational Photography