ESRGAN과 Semantic Soft Segmentation을 이용한 객체 분할의 성능 개선

윤동식, 곽노윤
백석대학교 ICT학부
kevinds1106@naver.com, nykwak@bu.ac.kr

Performance Improvement of Object Segmentation Using ESRGAN and Semantic Soft Segmentation

DongSik Yoon, Noyoon Kwak
Division of ICT, Baekseok University

요 약

본 논문은 ESRGAN(Enhanced Super Resolution GAN)과 Semantic Soft Segmentation을 이용한 객체 분할의 성능 개선에 관한 것이다. 본 논문의 연구진이 이미 제안한 Mask R-CNN과 Semantic Soft Segmentation을 이용한 객체 분할 방법은 전반적으로 객체 분할 성능이 양호한 반면, 객체의 크기가 상대적으로 작으면 분할 성능이 저조해지는 문제점이 있었다. 본 논문은 이러한 문제점을 해결하기 위한 것으로, Mask R-CNN을 통해 검출된 객체의 크기가 일정 기준치 이하인 경우, ESRGAN을 통해 초해상화를 수행한 후, Semantic Soft Segmentation을 수행함으로써 소형 객체의 분할 성능을 개선함에 그 목적이 있다. 제안된 방법에 따르면, 기존의 방법에 비해 크기가 작은 객체의 분할 특성을 좀 더 효과적으로 개선할 수 있음을 확인할 수 있었다.

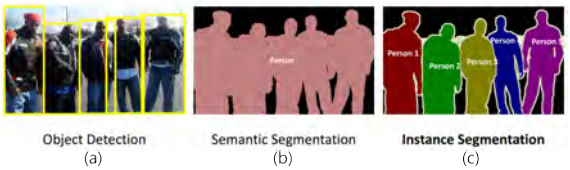
1. 서론

널리 알려진 Mask R-CNN[1]은 객체 검출(Object Detection)과 인스턴스 분할(Instance Segmentation) 두 측면에서 유용한 모델인데, 객체를 검출하는 성능은 우수한 편이지만 객체를 분할하는 성능은 기대에 미치지 못하는 경우가 많다. 이러한 점을 보완하기 위해 본 논문의 연구진은 Mask R-CNN과 Semantic Soft Segmentation[2]을 사용한 객체 분할 방법[3]을 기 제안한 바 있다. 하지만 이 방법에 따르면 Mask R-CNN을 통해 검출된 객체의 크기가 일정 이하인 경우에는 Semantic Soft Segmentation의 성능이 미진해 오히려 분할결과가 이전보다 나빠지는 문제점이 있다. 본 논문은 이러한 문제를 해결하기 위해 Mask R-CNN을 통해 검출한 객체의 크기가 일정 이하인 경우 ESRGAN (Enhanced Super Resolution GAN)[4]을 사용하여 초해상화를 수행한 후 Sementic Soft Segmentation을 수행함으로써 그 성능을 개선함에 목적이 있다.

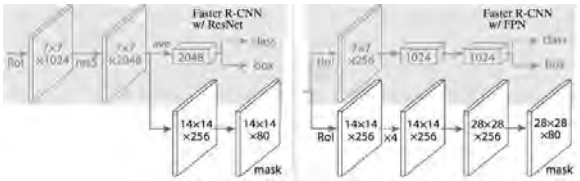
2. 관련 연구

2.1 Mask R-CNN을 이용한 객체 검출 및 분할

본 논문에서 객체 검출(Object Detection) 및 Instance Segmentation을 위해 채택한 Mask R-CNN[1]은 Faster R-CNN[5]에서 가능한 경계박스 단위의 분할과 FCN(Fully Convolutional Network)[6]에서 가능한 Semantic Segmentation을 결합한 것으로, 경계박스 단위로 전경 객체를 별도로 구별해 분할하는 알고리즘이다. Faster R-CNN이 객체 검출만을 위한 모델이라면, Mask R-CNN은 이를 확장하여 검출된 경계박스 내 마스크를 학습시키는 모델이다. 이 모델은 객체를 분할하기 위해 물체를 분류하는 브랜치와 경계박스를 회귀(regression)하는 브랜치에 객체 마스크를 예측하는 브랜치를 추가한 것이다. 즉 Mask R-CNN은 Faster R-CNN에 각 픽셀이 객체에 해당하는지 아닌지를 예측하는 추가적인 마스크 브랜치를 삽입해 경계박스 내 각각의 픽셀이 그 객체의 일부인지를 판별하는 이진 마스크를 구한다.



(그림 1) (a) Faster R-CNN에서의 객체 검출, (b) FCN에서의 Semantic Segmentation, (c) Mask R-CNN에서의 Instance Segmentation[5][6][1]



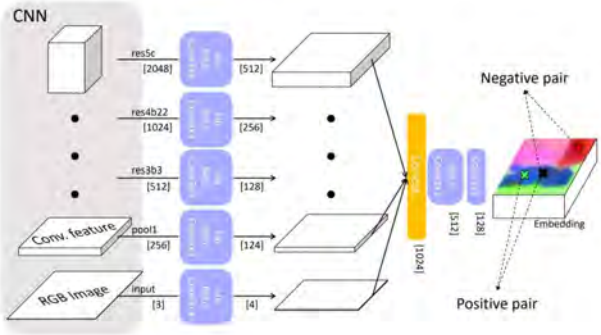
(그림 2) Faster R-CNN(ResNet/FPN)에 마스크 브랜치를 추가한 구조

통상, Mask R-CNN에서는 정확한 픽셀 위치가 필요하기 때문에 양선형 보간(bilinear interpolation)을 사용하는 RoIAlign을 추가적으로 채택함으로써 오정합 현상을 방지하고 각 특징맵과 원본 이미지의 해당 영역이 더 잘 정합되도록 한다.

2.2 Semantic Soft Segmentation

2.2.1 특징벡터 추출

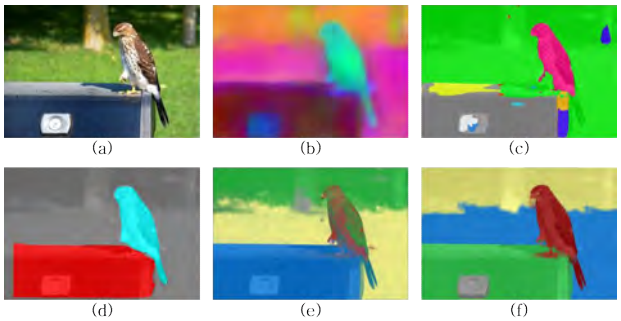
본 논문에서 의미론적 분할을 위해 사용한 Semantic Soft Segmentation[2]은 DeepLab ResNet-101 기반[7]의 네트워크를 사용하여 128차원의 벡터를 추출한다. 그림 3에서와 같이 ResNet-101의 'input', 'pool1', 'res3b3', 'res4b22', 'res5c'층을 추출해 나온 특징들을 3×3 합성곱을 취한 후 양선형 보간법을 사용하여 원본 이미지의 해상도로 업샘플링하고 이를 모두 결합하여 ReLu에 통과시킨다. 이때 중간 특징의 차원은 {3, 256, 512, 1024, 2048}인데 이를 {4, 124, 128, 256, 512}로 각각 압축하여 총 1024의 차원이 되게 한다. 그 후 두 번의 1×1 합성곱층을 통해 1024의 특징 차원을 512차원, 128차원으로 단계적으로 감소시킨다. 이 네트워크의 훈련에는 COCO Stuff dataset이 사용되었다.



(그림 3) Semantic Soft Segmentation의 특징 벡터 추출 네트워크의 구조

2.2.2 변형된 Spectral Matting을 이용한 의미론적 분할

그래프 이론 기반의 라플라시안 행렬(Laplacian matrix)을 사용하는 Spectral Matting[8]은 작은 패치들을 저수준 통계 특성(low-level statistics)만으로 다루기 때문에 객체를 판별하는 능력이 제한되는 단점이 있다. 이러한 단점을 보완하기 위해 Semantic Soft Segmentation은 라플라시안 행렬에 의미론적 특징을 혼합하고 장면 객체(scene objects)와 같은 고수준 개념(high-level concepts)을 포착해 이미지를 폭넓은 시점에서 판별하게 한다.



(그림 4) (a) 원본 이미지, (b) PCA기법을 사용해 128차원에서 추출한 특징벡터, (c) PSPNet을 이용한 분할 결과, (d) Mask R-CNN을 이용한 분할 결과, (e) Spectral Matting을 이용한 분할 결과, (f) Semantic Soft Segmentation을 이용한 분할 결과[9][1][8][2]

Semantic Soft Segmentation은 색상 기반의 원거리 상호작용을 하는 저수준의 유사도항을 비국부 색상 유사도(nonlocal color affinity)로 정의하였다. Semantic Soft Segmentation은 과분할된 이미지 기반의 가이디드 샘플링(guided sampling)을 진행하는데, 이미지 크기의 20% 이내에 해당하는 반지름을 가진 2,500개의 슈퍼픽셀[10]을 생성하고 각 슈퍼픽셀들 간의 유사도를 측정한다. 이 방법은 특징을 대표하기 충분한 각 슈퍼픽셀들을 하나의 샘플로 사용하며, 큰 반경을 사용하기 때문에 연결되지 않은 영역끼리의 연결도 가능하게 하는 장점이 있다. 추가적인 정보 없이 비국부 색상 유사도로 확장된 공간상에서 상호작용을 통한 이미지 분할을 진행할 시, 색상이 유사한 다른 객체들을 하나로 합치는 문제가 발생할 수 있다. Semantic Soft Segmentation은 의미론적으로 유사한 영역들이 하나의 덩어리로 분할되도록 세분화하기 위해 의미론적 유사도항을 추가하였다. 이는 같은 장면 객체들의 픽셀들은 서로 같이 분류되도록 도와주면서 서로 다른 객체끼리는 분리되도록 도와준다.

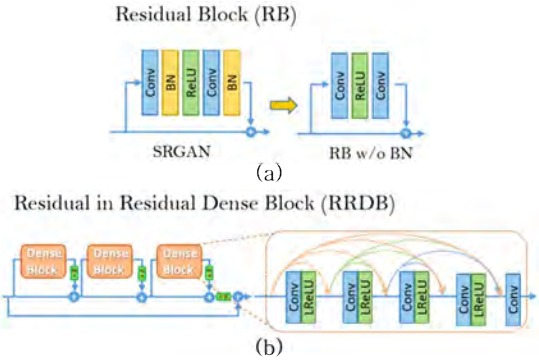
Semantic Soft Segmentation은 DeepLab ResNet-101 기반의 네트워크를 사용하여 나온 128차원의 벡터를 PCA기법을 통해 3차원으로 줄이고 기존의 Spectral Matting 연산 시 사용하는 라플라시안 행렬에 새로 만든 비국부 색상 유사도와 의미론적 유사도를 추가해 의미론적 분할을 진행하여 더 부드러운 영상 분할을 수행한다.

2.3 ESRGAN(Enhanced Super Resolution GAN)

기존의 SRGAN(Super Resolution GAN)은 한 장의 이미지만으로 초해상화를 수행하면서 현실적인 질감의 결과를 생성한다. 하지만 부분적으로 시각적 거부감이 초래되는 바, ESRGAN[4]은 세 가지 핵심 요소를 추가해 화질을 개선한다.

2.3.1 Residual-in-Residual Dense Block(RRDB)

본 논문에서 이미지 초해상화를 위해 사용된 ESRGAN은 기존의 SRGAN에서 화질을 향상시키기 위해 생성자의 구조에서 다음과 같은 두 가지의 보완 사항을 추가하였다. 그 하나는 모든 BN(Batch Normalization)층을 삭제한 것이고, 또 다른 하나는 그림 5(b)와 같이 제안된 Residual-in-Residual Dense Block(RRDB)을 사용하는 것이다.



(그림 5) (a) SRGAN의 Residual Blcok에서 BN층을 제거한 구조 (b) 제안된 Dense Block을 사용한 RRDB의 구조

BN층은 배치의 평균과 분산을 사용해서 특징들을 정규화하여 훈련하고 테스트 시에는 전체 데이터셋의 평균과 분산을 사용한다. 이 때문에 훈련과 테스트 데이터셋의 통계가 많이 다른 경우 BN층은 시각적 거부감을 생성하고 일반화하는 성능을 제한하는 경향이 있다. ESRGAN은 BN층을 제거함으로써 훈련을 안정화시키면서 일반화 성능을 높이고 계산 복잡도를 줄였다.

ESRGAN의 고차원 네트워크 구조는 SRGAN을 유지하면서 그림 5(b)와 같이 RRDB라는 새로운 기본 블록을 사용하였다. 더 많은 층과 연결이 있는 경우 성능이

더 잘나온다는 Residual Networks[11]의 실험을 기반으로 제안된 RRDB는 SRGAN의 Residual Block보다 더 깊고 복잡한 구조를 가진다. 추가로 ESRGAN은 향상된 구조를 통해 매우 깊은 네트워크를 학습시킬 수 있는 두 가지 기술을 추가로 활용한다. Residual Scaling은 학습의 불안정을 방지하기 위해 연산된 블록을 주경로(main path)와 결합하기 전에 0에서 1사이의 상수를 곱해서 블록을 축소한다. 또 Residual 네트워크 구조가 시작 파라미터의 분산이 작은 경우에 학습이 더 잘되는 것을 이용해 작은 초기값을 사용한다.

2.3.2 상대 판별자(Relativistic Discriminator)

ESRGAN은 생성자의 구조를 향상시키는 것뿐만 아니라 Relativistic GAN[12]을 기반으로 판별자도 강화한다. 기존의 SRGAN의 표준 판별자는 입력 이미지 x가 얼마나 진짜 같은지에 대한 가능성을 판별하는 것이었다면 ESRGAN의 판별자는 진짜 이미지 x_r이 생성한 가짜 이미지 x_f보다 얼마나 더 진짜 같은지를 예측하는 상대 평균 판별자(Relativistic average Discriminator)를 사용한다.

SRGAN의 표준 판별자는 $D(x)=\sigma(C(x))$의 식으로 나타낼 수 있다. 이때 σ는 시그모이드 함수이고 $C(x)$는 아직 변환되지 않은 판별자의 출력을 의미한다. ESRGAN에서 사용한 상대 평균 판별자의 식은 $D_{Ra}(x_r,x_f)=\sigma(C(x_r)-\mathbb{E}_{x_f}[C(x_f)])$로 나타낼 수 있다. 여기서 $\mathbb{E}_{x_f}[C(x_f)]$는 미니 배치에 있는 모든 가짜 데이터의 평균을 의미한다. 이를 사용한 판별자의 손실함수는 식 (1)과 같으며 생성자의 손실함수는 식 (2)와 같이 판별자의 손실함수와 대칭적인 형태를 보인다.

$$L_D^{Ra}=-\mathbb{E}_{x_r}[\log(D_{Ra}(x_r,x_f))]-\mathbb{E}_{x_f}[\log(1-D_{Ra}(x_f,x_r))] \quad (1)$$

$$L_G^{Ra}=-\mathbb{E}_{x_r}[\log(1-D_{Ra}(x_r,x_f))]-\mathbb{E}_{x_f}[\log(D_{Ra}(x_f,x_r))] \quad (2)$$

식 (1) 및 (2)에서 $x_f=G(x_i)$이며 x_i는 입력 LR(Low Resolution) 이미지를 의미한다. 식 (2)에서와 같이 생성자의 손실함수는 x_r와 x_f 모두를 포함하기 때문에 SRGAN과 달리 생성된 이미지와 원본 이미지 모두 경쟁적 학습에 영향을 미치는 것을 볼 수 있다.

2.3.3 지각 손실(Perceptual Loss)

ESRGAN은 좀 더 효과적인 지각 손실(perceptual loss)을 이용한다. 보통의 지각 손실은 사전 학습된 네트워크의 활성화 층을 통과한 특징을 최소화하는 것으로 정의한다. 이와 달리 ESRGAN에서는 활성화 층을 통과하기 전의 특징을 사용하는 것으로 두 가지 문제를 개선한다. 네트워크가 매우 깊은 경우 활성화되는 특징들이 매우 드물게 나타나기 때문에 좋지 못한 결과를 초래하는 문제와 활성화 층을 통과한 특징을 사용하는 경우 원본 이미지와 비교하여 재구성된 밝기가 일관되지 않은 문제이다.

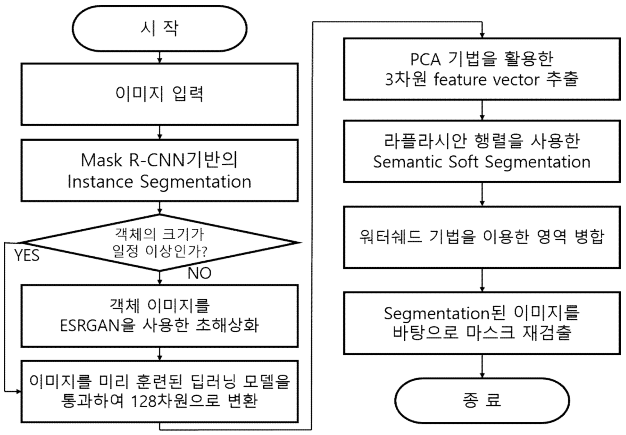
결론적으로 ESRGAN의 생성자 손실 함수는 식(3)과 같이 정의할 수 있다.

$$L_G=L_{percep}+\lambda L_G^{Ra}+\eta L_1 \quad (3)$$

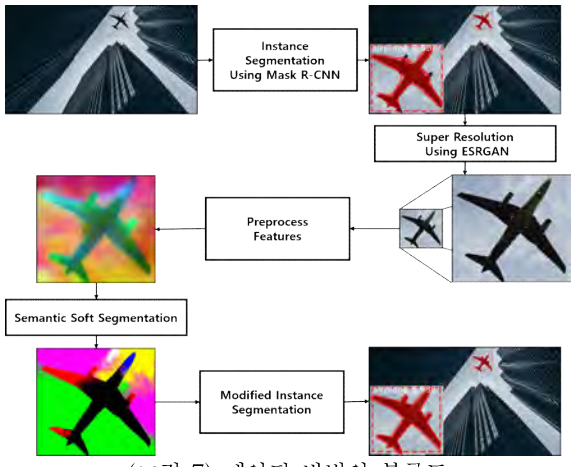
식 (3)에서 $L_1=\mathbb{E}_{x_i}\|G(x_i)-y\|_1$은 콘텐트 손실로, 생성된 이미지 $G(x_i)$와 원본 이미지 y의 차이를 구한 것이다. λ와 η는 다른 손실과의 균형을 맞추기 위한 계수이다.

3. 제안된 객체 분할 방법

본 논문은 Mask R-CNN으로 객체를 검출하고 Semantic Soft Segmentation을 통해 의미론적 분할을 수행한다. 이후 상기 검출 객체와 상기 분할 결과를 대응시켜 최종적으로 객체를 분할한다. 이때 Mask R-CNN의 검출 객체의 크기가 일정 기준치 이하라면 ESRGAN을 활용해 해상도를 높인 이미지를 생성하여 Semantic Soft Segmentation을 수행한다. 결과적으로 Mask R-CNN의 검출 객체의 크기가 작은 경우에 Semantic Soft Segmentation의 분할 성능을 개선하는 효과가 있다. 그림 6은 제안된 방법의 순서도를 나타낸 것이고, 그림 7은 제안된 객체 분할 블록도를 나타낸 것이다.



(그림 6) 제안된 방법의 순서도



(그림 7) 제안된 방법의 블록도

3.1 객체 검출 및 초해상화

- 객체 검출 및 분할: Mask R-CNN 모델을 활용하여 입력 이미지에서 객체를 검출하고 분할한다. Mask R-CNN 모델은 그림 8(a)와 같이 각 객체의 경계박스 정보와 객체 분할된 마스크 정보, 객체 종류(class_ids), 정확도를 반환한다.
- 초해상화: Mask R-CNN을 활용하여 검출된 객체 중에서 크기가 일정 기준치 이하인 경우, ESRGAN을 사용해 그림 8(b)와 같이 소형 객체 이미지를 대상으로 4배 초해상화를 수행한다.

3.2 Semantic Soft Segmentation의 영상 분할

- 특징 벡터 추출: ResNet-101 기반의 사전 훈련된 네트워크를 사용하여 128차원의 벡터를 추출한다. 그 후 PCA 기법을 사용해 그림 8(c)와 같이 3차원의 특징 벡터를 추출한다.
- Semantic Soft Segmentation 수행: PCA 기법을 활용해 3차원으로 만든 특징 벡터와 원본 이미지를 사용해 그림 8(d)와 같이 Semantic Soft Segmentation을 수행한다.

(a) (b) (c) (d)

(그림 8) (a) 객체 검출을 완료한 이미지, (b) ESRGAN을 활용해 객체를 초해상화한 이미지, (c) PCA기법을 통해 특징 벡터를 추출한 이미지, (d) Semantic Soft Segmentation을 수행한 이미지

3.3 워터쉐드 기법을 활용한 영역 분할

- 워터쉐드 마커 설정: 워터쉐드 기법을 통해 의미론적 분할 결과의 보정 및 영역 병합을 수행하기 위해 확실한 배경인 그림 9(a)와 확실한 전경인 그림 9(b)를 각각의 마커들로 설정한다.
- 워터쉐드 분할 수행: 특징 벡터인 그림 8(c)와 마커로 설정한 그림 9(c)를 참조하여 그림 9(d)와 같은 워터쉐드 분할을 수행한다.

3.4 객체 분할

- 객체 마커 탐색: 워터쉐드 기법으로 분할된 여러 마커들 중 의미론적 분할 결과를 기반으로 검출된 객체에 해당하는 마커들을 찾는다.
- 객체 분할: 객체에 해당되는 마커들을 Mask R-CNN의 마스크와 교체하여 객체 분할을 완료한다.

(a) (b) (c) (d)

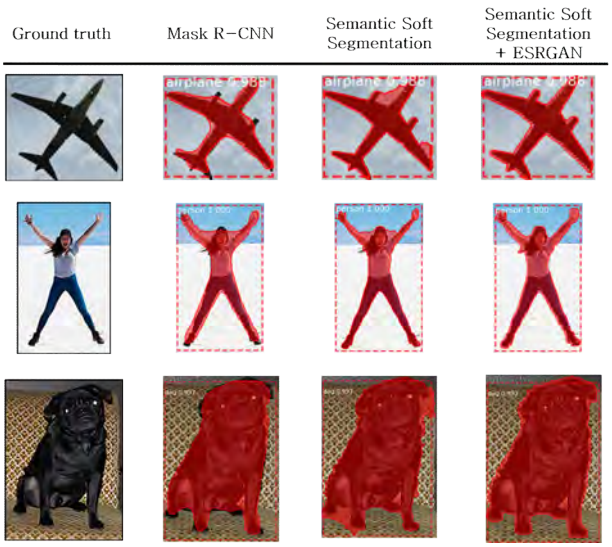
(그림 9) 의미론적 분할 결과 확실한 전·배경에 마커를 설정한 이미지(c), 그림 8(c)와 그림 9(c)를 참조하여 워터쉐드를 수행한 이미지 (d)

4. 시뮬레이션 결과 및 고찰

본 논문에서는 Jupyter Notebook과 NVIDIA GTX 1070Ti GPU 환경에서 사전훈련된 Mask R-CNN과 Semantic Soft Segmentation 및 ESRGAN 모델을 사용하여 컴퓨터 시뮬레이션을 수행하였다. 그림 10은 기존의 방법과 제안된 방법의 결과를 비교한 것으로 전경과 배경 구분이 뚜렷함에도 불구하고 객체 분할 결과가 만족스럽지 못하다. 반면에 제안된 방법은 상대적으로 우수한 객체 분할 성능을 제공함을 시각적으로 확인할 수 있었다.

5. 결론

제안된 방법은 Mask R-CNN과 Semantic Soft Segmentation을 이용한 객체 분할 시 검출된 객체의 크기가 작은 경우, ESRGAN을 이용한 초해상화를 통해 Semantic Soft Segmentation의 의미론적 분할 성능을 개선한 것이다. 제안된 방법은 검출된 객체의 크기가 작더라도 우수한 분할 성능을 제공함을 확인할 수 있었다. 이를 감안할 때 추가적인 성능 개선이 이뤄진다면 고품질의 영상 편집, 객체 인식 및 트래킹 영상 이해 등이 가능할 것으로 기대된다. 또 제안된 방법을 통해 기존의 COCO 데이터셋과 같이 객체 분류 모델을 학습시키기 위해 사람의 손으로 일일이 생성해주던 데이터셋을 수작업을 거치지 않고 자동으로 생성해 비용과 시간을 절감하고 좀 더 세밀한 데이터셋 기반의 학습으로 모델의 성능 향상을 기대할 수 있을 것이다.



(그림 10) 기존의 객체 분할 방법과 제안된 객체 분할 방법

참고문헌

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. IEEE International Conference on Computer Vision*, pp. 2980-2988. 2017.

[2] Y. Aksoy, T.-H. Oh, S. Paris, M. Pollefeys, and W. Matusik, "Semantic Soft Segmentation," *ACM Trans. Graph.*, 2018.

[3] 윤동식, 곽노윤, "MaskR-CNN과 Semantic Soft Segmentation을 이용한 객체 분할", 2020년도 한국통신학회 동계종합학술발표회 논문집, pp. 872-873, 2020. 2.

[4] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C.-C Loy, Y. Qiao and X. Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," *Proc. ECCV*. 2018.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," *in Neural Information Processing Systems*, 2015.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol 40, 2017.

[8] A. Levin, A. Rav-Acha, and D. Lischinski, "Spectral Matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1699 - 1712, Oct. 2008.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. "Pyramid Scene Parsing Network," *Proc. CVPR* 2017.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274-2281, Nov. 2012.

[11] Lim, B, Son, S., Kim, H, Nah, S., Lee, KM, "Enhanced deep residual networks for single image super-resolution," *Proc. CVPR* 2017.

[12] Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.