논문 2021-16-27

딥 러닝 기반의 팬옵틱 분할 기법 분석

(Survey on Deep Learning-based Panoptic Segmentation Methods)

권 정 은, 조 성 인*
(Jung Eun Kwon, Sung In Cho)

Abstract : Panoptic segmentation, which is now widely used in computer vision such as medical image analysis, and autonomous driving, helps understanding an image with holistic view. It identifies each pixel by assigning a unique class ID, and an instance ID. Specifically, it can classify 'thing' from 'stuff', and provide pixel-wise results of semantic prediction and object detection. As a result, it can solve both semantic segmentation and instance segmentation tasks through a unified single model, producing two different contexts for two segmentation tasks.
Semantic segmentation task focuses on how to obtain multi-scale features from large receptive field, without losing low-level features. On the other hand, instance segmentation task focuses on how to separate 'thing' from 'stuff' and how to produce the representation of detected objects. With the advances of both segmentation techniques, several panoptic segmentation models have been proposed. Many researchers try to solve discrepancy problems between results of two segmentation branches that can be caused on the boundary of the object. In this survey paper, we will introduce the concept of panoptic segmentation, categorize the existing method into two representative methods and explain how it is operated on two methods: top-down method and bottom-up method. Then, we will analyze the performance of various methods with experimental results.

Keywords : Panoptic segmentation, Thing, Stuff, Top-down method, Bottom-up method

Ⅰ. 서 론

딥러닝은 사람의 사고능력을 모방한 인공 신경망 기술로, 현재 널리 주목받고 있는 기술이다. 컴퓨터 비전 분야에서는 딥러닝을 활용한 이미지 분류 (image classification), 객체 탐지 (object detection), 객체 분할 (object segmentation) 등의 문제 해결 방법들이 제안되고 있다.

이 중 분할 기법은 그림 1과 같이 크게 세 가지로 분류될 수 있다. 이미지 내 모든 픽셀의 레이블을 예측하는 의미론적 분할 기법 (semantic segmentation) [1]과, 각 객체를 구분하여 개별적인 객체 아이디를 부여하는 객체 분할 기법 (instance segmentation) [2]이 있다. 최근에는 객체뿐만 아니라 배경 정보까지 함께 검출하기 위해 두 가지 분할 기법을 결합한 팬옵틱 분할 기법 (panoptic segmentation) [3]에 대한 연구가 진행되고 있다. 이는 픽셀 단위로 객체의 클래스를 예측할 뿐만 아니라, 개별적인 객체 아이디도 부여하는 분할 기법이다.

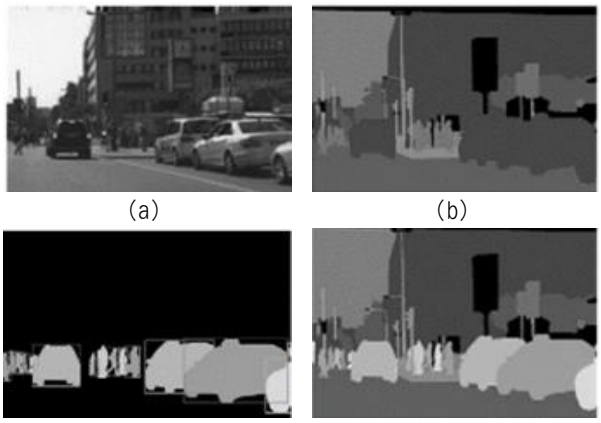

그림 1. 분할 기법 비교. (a) 입력 영상, (b) Semantic segmentation 결과, (c) Instance segmentation 결과, (d) Panoptic segmentation 결과.
Fig. 1. Comparison of segmentation methods. (a) Input image, (b) Result of semantic segmentation, (c) Result of instance segmentation, (d) Result of panoptic segmentation.

팬옵틱 분할 기법에서는 객체의 정보를 불가산 객체 (stuff) 와 가산 객체 (thing)로 구분하여 분할을 적용한다. 불가산 객체는 무정형의 불가산 영역을 의미하며, 대표적으로 하늘, 도로 등이 이에 해당한다. 가산 객체는 객체 아이

*Corresponding Author (csi2267@dongguk.edu)
Received: Aug. 20, 2021, Revised: Oct. 1, 2021, Accepted: Oct. 6, 2021
J.E. Kwon: Dongguk University (Integrated MS program)
S.I. Cho: Dongguk University (Prof.)
※ This research was supported by Samsung Electronics, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1C1C1009662), and Basic Science Research Program through the NRF funded by Ministry of Education (No. NRF-2020X1A3A1093880).

ⓒ IEMEK J. Embed. Sys. Appl. 2021 Oct. 16(5) 209-214
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2021.16.5.209

디가 부여되는 객체를 의미하며, 차, 트럭, 사람 등이 이에 해당한다. 의미론적 분할 기법으로는 이미지의 모든 픽셀에 대하여 불가산 객체와 가산 객체에 대한 구분 없이 픽셀단위의 클래스 예측을 한다. 객체 분할 기법으로는 가산 객체로 분류된 클래스에 대하여 객체 아이디, 클래스, 마스크에 대한 예측을 한다. 그림 1에서는 동일한 입력 영상에 대한 의미론적 분할 기법, 객체 분할 기법, 팬옵틱 분할 기법 각각의 결과를 보여준다.

팬옵틱 분할 기법은 크게 top-down method [4, 5]와 bottom-up method [6-8]로 구분된다. Top-down method는 객체 분할 기법과 의미론적 분할 기법을 병렬적으로 수행하는 방법이다. 반면, bottom-up method는 의미론적 분할 기법을 한 후, 해당 정보를 이용하여 객체 분할 기법을 하는 순차적인 방법이다. 다음 장에서는 각 top-down method와 bottom-up method에 해당하는 대표적인 모델들의 구조와 주요 동작원리에 대해서 분석한다.

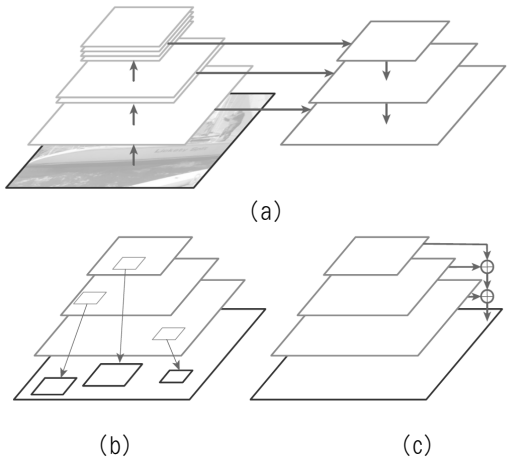

그림 2. Panoptic-FPN [4]의 구조. (a) Feature Pyramid Network (FPN), (b) 객체 분할 기법 가지, (c) 의미론적 분할 기법 가지.
Fig. 2. Structure of Panoptic-FPN. (a) Feature Pyramid Network (FPN), (b) Instance segmentation branch, (c) Semantic segmentation branch.

Ⅱ. 기존 연구 분석

1. Top-down methods

Top-down method [4, 5]는 객체 분할 기법과 의미론적 분할 기법을 병렬적으로 수행하는 방법이다. 이 방법에서는 크게 객체 분할 기법 가지와 의미론적 분할 기법 가지가 존재한다. 객체 분할 기법 가지는 region of interest (ROI)를 지정하는 단계와, 지정한 영역을 바탕으로 가산 객체를 검출하는 단계를 가지는 two-stage 구조다. 의미론적 분할 기법 가지에서는 전체 이미지의 픽셀 단위 클래스를 예측한다. 이후, 두 가지에서의 결과를 비교 및 대조하여 픽셀에 대한 클래스 아이디와 객체 아이디를 예측한다. 본 장에서는 대표적인 top-down 구조 모델인 Panoptic Feature Pyramid Network (Panoptic-FPN) [4]과 A Unified Panoptic Segmentation Network (UPSNet) [5]를 소개한다.

Panoptic-FPN [4]는 하나의 backbone에 두 가지: 객체 분할 기법 가지, 의미론적 분할 기법 가지를 추가하여 팬옵틱 분할 기법으로 확장한 구조이다. 각 구성 요소에 대한 구조는 그림 2에 표현되어 있다. Backbone으로 사용되는 Feature Pyramid Network (FPN) [9]는 객체 탐지에서 multi scale 정보를 효과적으로 사용하기 위해 제안된 방법이다.

객체 분할 기법 가지에서는 Mask R-CNN [2] 구조의 가지를 사용해, FPN backbone으로부터 추출된 특징 맵(feature map)으로부터 객체 분할 기법에 대한 logit인 객체의 위치, 클래스 아이디와 픽셀 단위 예측인 mask logit을 얻는다. 의미론적 분할 기법 가지에서는 FPN backbone으로부터 추출된 각 특징 맵의 크기를 업샘플링한 후 픽셀 단위로 더하여 하나의 특징 맵을 만든다. 이후, convolution과 이중 선형 보간법, softmax를 통해 픽셀 단위로 클래스 예측

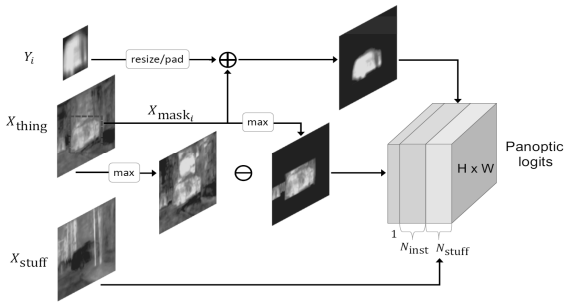

그림 3. UPSNet [5]의 panoptic segmentation head의 구조.
Fig. 3. Structure of panoptic segmentation head in UPSNet [5].

을 하는 semantic logit을 얻는다.

최종적으로 각 가지에서 도출한 semantic 클래스 아이디와 객체 아이디의 pair (p_i, z_i)로 팬옵틱 분할 결과를 만들어 낸다. 이 때 픽셀에서의 클래스 아이디를 p_i, 객체 아이디를 z_i라 한다. 객체 아이디와 클래스 아이디가 겹칠 경우 객체 아이디만을 사용한다. 객체 아이디가 부여된 픽셀일 경우 클래스 아이디를 'other'로 구분한다. 객체에 해당하지 않는 픽셀일 경우 객체 아이디를 'none'으로 구분한다. 객체 분할 기법 가지에서의 손실함수는 다음과 같다.

$$L_{instance\,branch} = L_c + L_b + L_m, \quad (1)$$

여기서, L_c은 classification loss를, L_b은 bounding box loss를, 그리고 L_m은 mask loss를 의미한다. 의미론적 분할 기법 가지에서는 교차 엔트로피 손실 함수 을 사용해 픽셀단위의 클래스를 예측한다. 최종 loss L은 각 가지에서의 loss를 정규화하여 계산한다.

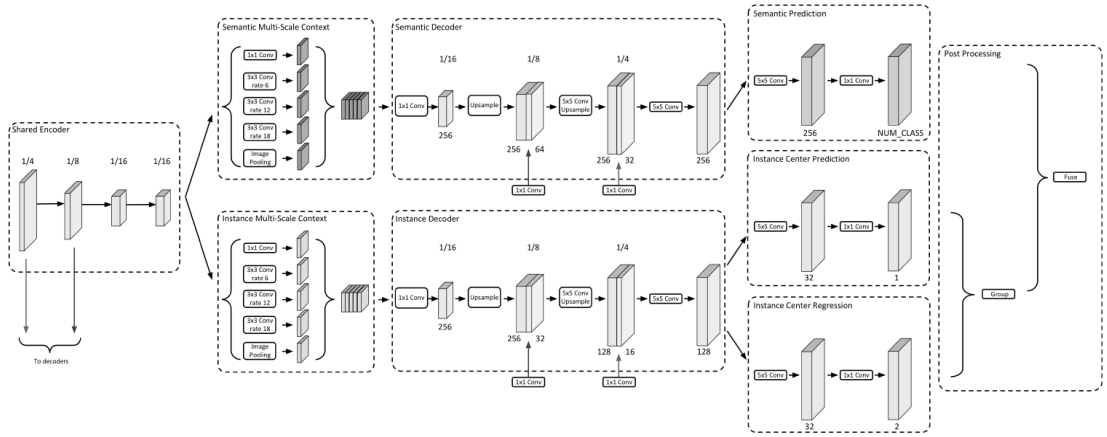

그림 4. Panoptic-DeepLab의 구조 [7].
Fig. 4. Structure of Panoptic-DeepLab [7].

UPSNet [5]는 Panoptic-FPN에 panoptic logit을 만드는 팬옵틱 분할 가지를 추가한 구조이다. panoptic segmentation head 부분의 구조는 그림 3과 같다. $(N_{stuff}+N_{inst}+1)\times H\times W$ 의 크기를 가지는 panoptic logit은 픽셀을 가산 객체 또는 불가산 객체로 분류할 뿐만 아니라, 검출되지 않은 가산 객체에 대해서는 unknown으로 분류한다. 여기서 N_{stuff}은 불가산 객체 클래스의 개수, N_{inst}은 객체에 해당하는 클래스의 개수이다. 객체 분할 기법 가지와 의미론적 분할 기법 가지에서의 각 결과를 하나의 panoptic logit으로 표현하기 위해 panoptic segmentation head에서 다음과 같은 과정을 수행한다. 먼저, 의미론적 분할 기법 가지에서의 클래스를 불가산 객체와 가산 객체로 나눈다. panoptic logit의 채널에서는 semantic logit의 불가산 객체 클래스 채널에 해당하는 logit을 사용한다. panoptic logit의 채널에서는 객체 분할 기법 가지의 결과와 semantic logit과 비교하여 픽셀단위 객체 아이디와 클래스 아이디를 부여한다. panoptic logit의 unknown 채널에서는 같은 객체에 대해 semantic logit과 instance segmentation logit이 예측한 객체 클래스를 비교하여 두 신뢰 값의 차를 사용한다. 위의 세 logit으로 panoptic logit을 채운 후, 채널방향으로 softmax하여 픽셀 단위 클래스 예측과 객체 아이디 예측이 가능하도록 한다.

UPSNet에서는 의미론적 분할 기법 가지 학습 시, 객체 분할에서 검출된 객체에 대해 가중치를 부여하기 위해 Region of Interest (ROI) loss를 제안한다. 이 두 가지 방법은 팬옵틱 분할 기법의 정확도를 기존보다 향상시켰다. 이후의 팬옵틱 가지 학습 시에는, 픽셀 단위 클래스 아이디와 객체 아이디 예측을 위해 교차 엔트로피 손실함수를 사용하였다.

2. Bottom-up methods

Bottom-up method [6-8]은 의미론적 분할 기법과 객체 분할 기법을 순차적으로 수행하는 방법이다. DeeperLab [6]은 클래스 정보가 주어지지 않은 객체를 탐지하기 위해서 객체의 무게중심 (instance center)을 검색하는 초기의 bottom-up 기법이다. 본 섹션에서는 DeeperLab을 기반으로 한 대표적 bottom-up 구조 모델인 Panoptic-DeepLab [7]과 Axial-DeepLab [8]의 구조와 주요 동작원리에 대해 분석한다.

먼저, Panoptic-DeepLab의 전체적인 구조는 그림 4와 같다. Spatial Pyramid Pooling (SPP)에 Atrous convolution을 적용한 ASPP 모듈을 backbone의 말단 블록에 적용하여 여러 비율에 따른 multi scale의 특징을 추출한다. 이러한 구조는 파라미터 수를 늘리지 않고도 한 픽셀이 볼 수 있는 수용 영역 (receptive field)을 넓힘으로써 조밀한 특징 맵 (feature map)을 만들어 큰 객체의 특징을 잘 검출할 수 있게 한다. 이후, decoder 모듈에서 컨텍스트를 저수준 특징 (low level feature)인 backbone 특징과 결합함으로써 업샘플링시 생길 수 있는 정보손실을 방지한다.

그림 4의 task specific head 중 하나인 semantic prediction head에서는 교차 엔트로피 손실 함수를 사용해 의미론적 분할 기법을 수행한다. Instance center prediction head에서 객체의 무게중심을 학습하기 위해 사용한 평균 제곱 오차 손실함수는 다음과 같다.

$$\begin{aligned} L_{instance\,center} &= MSE(y_{center}, t_{heatmap}) \\ &= \frac{1}{N}\Sigma_{(i=1)}^{N}(y_{center}-t_{heatmap})^2, \end{aligned} \tag{2}$$

(2)에서 y_{center}은 예측된 무게중심을 의미하고, $t_{heatmap}$은 무게중심의 ground truth에 gaussian blur를 적용하여 만든 히트맵을 의미한다. Instance center regression head에서 instance regression을 학습하기 위해 사용한 손실함수는 다음과 같다.

$$\begin{aligned} L_{instance\,regression} &= L_1(y_{center}, t_{thing}) \\ &= \Sigma_{(i=1)}^{n}|y_{center}-t_{thing}|, \end{aligned} \tag{3}$$

(3)에서 y_{center}은 예측된 객체의 무게중심을 의미하고, t_{thing}은 가산 객체로 예측된 픽셀을 의미한다. 도출한 객체의 무게중심과 편차를 이용해 픽셀에 객체 아이디를 부여하여 군집화 함으로써 객체 분할을 수행한다.

$$k_{(i,j)} = \min \| C_k - ((i,j) + O(i,j)) \|^2 , \quad (4)$$

(4)에서 $k_{(i,j)}$은 (i,j)에 위치한 픽셀에서 예측한 객체 아이디이다. C_k은 픽셀 (i,j)에 위치하는 객체의 무게중심을, $O(i,j)$은 instance center regression head에서 예측한 편차를 의미한다. 이후, 두 분할 결과를 융합하여 최종적으로 팬옵틱 분할 결과를 예측한다.

Axial-DeepLab은 Panoptic-DeepLab과 구조 및 동작 원리는 동일하나, 수용영역을 넓히기 위한 방법으로 ASPP 모듈 대신 axial-attention [10]을 채택한다. Axial-attention은 기존 convolution과는 달리, 가중치와 convolution이 입력영상의 일부 영역에만 의존해 수행되는 self-attention [11]을 변형해 사용하는 방법이다. Axial-attention은 self-attention을 2차원이 아닌 각 폭, 너비 (height, width)에 해당하는 1차원 축에 적용함으로써 기존보다 낮은 복잡도로 팬옵틱 분할 기법의 정확도를 향상시켰다. 또한 적은 비용으로도 큰 수용 영역에서 관련된 컨텍스트 위치에 더 민감하게 반응할 수 있도록 하였다.

Ⅲ. 실험 결과

본 장에서는 앞서 소개한 top-down method [4, 5], bottom-up method [6-8]의 팬옵틱 분할 기법의 성능을 비교한다. 팬옵틱 분할의 정확도를 나타내는 지표인 PQ는 수식 (5)로 계산된다.

$$PQ = \frac{\sum_{(p,q)\in TP} IoU(p,q)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}, \quad (5)$$

여기서, PQ는 panoptic quality의 약자로, 분할 네트워크의 분할 성능을 평가하는 척도이며, 각 클래스별로 독립적으로 계산된다. PQ metrics는 옳게 검출한 (TP) 부분에 대한 mIoU 지표를 나타내는 SQ (segmentation quality)와 F1 score를 나타내는 RQ (recognition quality)로 분해할 수 있다. 틀린 검출 (FP)이거나, 검출되지 않아야 할 것이 검출된 경우 (FN) 해당 부분에 대해 정확도가 떨어지게 된다. 여기서 TP, FP, FN은 각각에 해당하는 객체의 개수를 의미하며 절댓값을 취해서 양수임을 표현해준 것이다.

팬옵틱 분할 기법의 결과로 나온 객체 탐지에 대한 평가지표로 AP (average precision)도 사용한다. 객체 탐지의 정확도를 예측 결과를 기준으로 판단하는 정밀도 (precision)와 ground truth를 기준으로 판단하는 재현율 (recall)을 종합하여 객체 탐지 알고리즘을 평가하는 지표로 사용되는 것이 AP이다.

표 1, 표 2를 이용해 cityscapes test dataset [11]과 cocotest dataset [12]에 대한 정확도를 두 지표로 판단하고 비교했다.

표 1은 cityscapes test dataset에서의 실험 결과의 PQ, AP을 보여준다. 해당 데이터 셋에서는 top-down method보다 bottom-up method에서 더 높은 PQ을 보였으며, 그 중에

표 1. Cityscapes 테스트 데이터셋 [11].
Table 1. Cityscapes test set [11].

Method	Model	Backbone	PQ	PQ_{th}	PQ_{st}	AP
Top-down	Panoptic-FPN	FPN	58.1	52.0	62.5	33.0
	UPSNet	ResNet-101-FPN	59.3	54.6	62.7	33.3
Bottom-up	Panoptic-DeepLab	XCeption-71	62.3	-	-	34.6
	Axial-DeepLab	Axial-ResNet-XL	62.8	-	-	34.0

표 2. COCO 테스트 데이터셋 [12].
Table 2. COCO test set [12].

Method	Model	Backbone	PQ	PQ_{th}	PQ_{st}
Top-down	Panoptic-FPN	FPN	40.9	48.3	29.7
	UPSNet	ResNet-101-FPN	46.6	53.2	36.7
Bottom-up	Panoptic-DeepLab	XCeption-71	41.4	45.1	35.9
	Axial-DeepLab	Axial-ResNet-XL	44.2	49.2	36.8

서도 Axial-DeepLab이 62.8%로 가장 높은 PQ을 보였다. 검출한 객체의 클래스를 가산객체와 불가산 객체로 나누어 각각의 정확도인 PQ_{th}와 PQ_{st}을 비교한 결과, bottom-up method보다 top-down method에서 62.5%, 62.7%로 더 높은 PQ 성능을 얻었다. 또한 Panoptic-DeepLab에서 34.6%로 가장 높은 AP을 얻었다.

표 2는 coco test dataset에서의 실험 결과의 PQ을 보여준다. 해당 데이터 셋에서는 Axial-DeepLab이 44.2%로 가장 높은 PQ을 보였다. 또한, top-down method와 bottom-up method에서 PQ_{st}과 PQ_{th}을 비교한 결과, PQ_{th}가 더 높은 결과를 보였다.

Ⅳ. 결 론

본 논문에서는 이미지 픽셀을 가산 객체와 불가산 객체로 분류하고 분할하는 팬옵틱 분할 기법의 top-down method와 bottom-up method를 서술하고, 각 데이터 셋에 대한 분할 성능을 비교했다.

Top-down method에서는 주어진 가산 객체의 정보를 학습하기 위해 기존의 객체 분할 기법을 이용해 객체의 클래스 아이디, 바운딩 박스, 마스크 logit을 추출하여 사용했다. 추출한 객체의 정보를 의미론적 분할 기법 가지가 예측한 클래스 아이디와 비교하여 팬옵틱 분할 기법의 결과를 추론하는 것이 top-down method의 핵심 과정이다.

Bottom-up method에서는 클래스 아이디가 주어지지 않은 가산 객체를 학습하기 위해 의미론적 분할 기법 이후, 주어진 가산 객체를 무게중심으로 표현하고, 선형회귀를 통해 해당 객체의 크기를 예측했다. 이처럼, 의미론적 분할 기법의 결과로부터 이미지의 픽셀들을 하나의 객체로 군집화 하는 것이 bottom-up method의 핵심 과정이라 할 수 있다.

팬옵틱 분할 기법은 이미지 내 픽셀들을 가산 객체와 불가산 객체로 나누어 이미지를 전체적으로 이해할 수 있게 하여, 기존의 두 분할 기법의 대체 가능성을 보였다. 본문에서 제시된 방법들은 객체 분할 기법과 의미론적 분할 기법에 의존적으로 작동했다. 따라서 두 분할 기법에 해당하는 모델의 정확도를 향상시키는 것은 더욱 중요하다. 또한, 두 분할기법에 의존하지 않고 하나의 가지만을 이용해 팬옵틱 분할 기법을 수행하는 연구의 필요성 또한 증가할 것이다.

References

[1] J. Long, E. Shelhamer, T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440, 2015.

[2] K. He, G. Gkioxari, P. Dollár, R. Girshick, “Mask r-cnn,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 2961-2969, 2017

[3] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, “Panoptic Segmentation,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9404-9413, 2019.

[4] A. Kirillov, R. Girshick, K. He, P. Dollár, “Panoptic Feature Pyramid Networks,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6399-6408, 2019.

[5] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, R. Urtasun “Upsnet: A Unified Panoptic Segmentation Network,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8818-8826, 2019.

[6] T. J. Yang, M. D. Collins, Y. Zhu, J. J. Hwang, T. Liu, X. Zhang, L. C. Chen, “Deeperlab: Single-shot Image Parser,” arXiv preprint arXiv:1902.05093, 2019.

[7] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, L. C. Chen,” Panoptic-deeplab: A Simple, Strong, and Fast Baseline for Bottom-up Panoptic Segmentation,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12475-12485, 2020

[8] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L. C. Chen, “Axial-deeplab: Stand-alone Axial-attention for Panoptic Segmentation,” In European Conference on Computer Vision, pp. 108-126, Springer, Cham, August 2020.

[9] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, “Feature Pyramid Networks for Object Detection,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117-2125, 2017.

[10] H. Zhao, J. Jia, V. Koltun, “Exploring Self-attention for Image Recognition,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10076-10085, 2020.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, B. Schiele, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Eecognition, pp. 3213-3223, 2016.

[12] A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie, “Coco-text: Dataset and Benchmark for Text Detection and Recognition in Natural Images,” arXiv preprint arXiv:1601.07140, 2016.

Jungeun Kwon (권 정 은)

2018~Multi-media Engineering, in Dongguk University, Seoul, Republic of Korea (Integrated MS program)

Field of Interests: Image Processing, Deep-learning, and Computer Vision

Email: kje_9912@dongguk.edu

Sungin Cho (조 성 인)

2010 Electronics Engineering from Sogang University, Seoul, Republic of Korea (B.S.)

2015 Electrical and Computer Engineering from Pohang University, Pohang, Republic of Korea (Ph.D.)

2015~2017 Senior Researcher with LG Display

2017~2019 Electronic Engineering from Daegu University, Gyeongsan, Republic of Korea (Assistant Prof.)

2019~ Multimedia Engineering, in Dongguk University (Assistant Prof.)

Field of Interests: Image Analysis and Enhancement, Video Processing, Computer Vision, and Deep-learning.

Email: csi2267@dongguk.edu