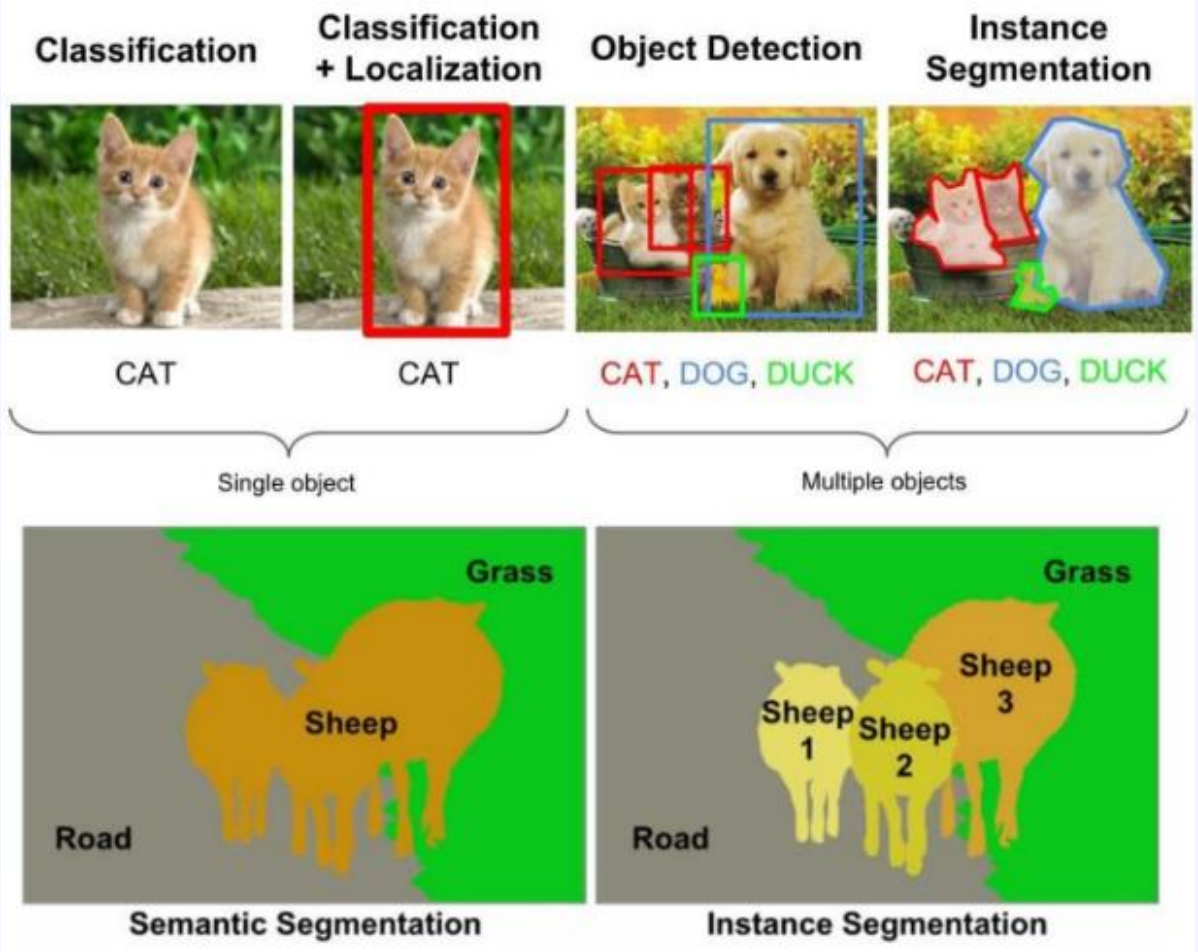
Ch5. HRNet v2 + OCR

Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation

Computer vision



Sematic segmentation

- Instance segmentation
  - 일반적으로 RGB image를 입력으로 사용
  - 관심 있는 객체를 찾고 찾은 객체에 대해 Pixel-wise classification

Instance segmentation

Semantic segmentation

Segmentation task

- Input
  - RGB color 이미지 (height X width X 3) 또는 흑백 (height X width X 1) 이미지
- Output
  - 각 필셀 별 어느 class 에 속하는지 나타내는 레이블을 나타낸 Segmentation Map

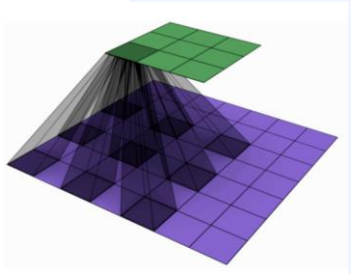
input

Segmented

Semantic Labels

Multi-scale context & Relation context

- Multi-scale context
  - 여러 스케일의 인풋들로 구성을해서 병렬적으로 처리하는 멀티스케일 컨텍스트라는 방법
- Relation context
  - Object region을 활용해서 pixel의 representation을 강화(augment)



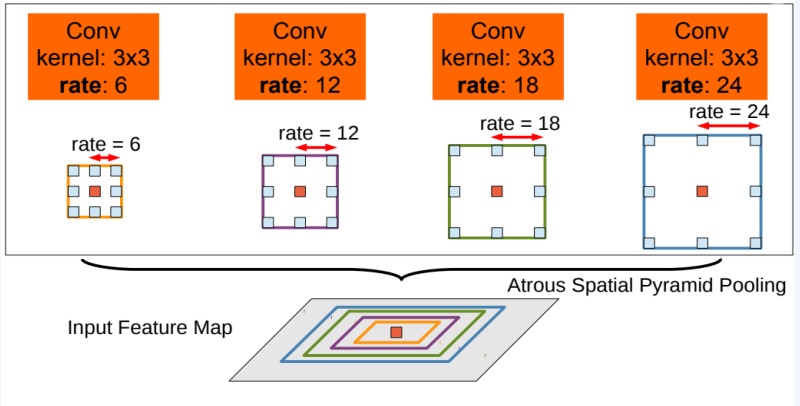
ASPP

OCR

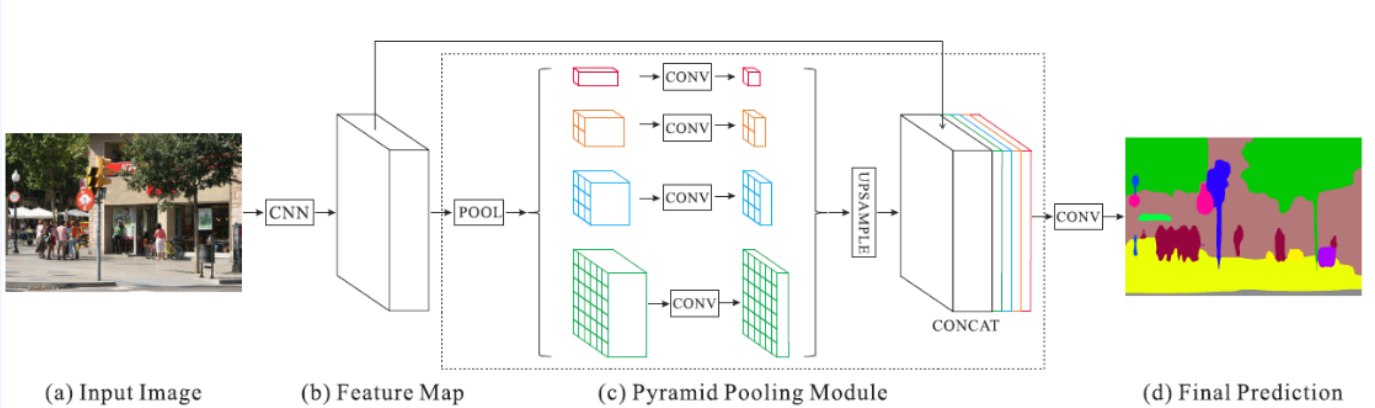
Multi-scale context

Multi-scale context

- ASPP(Atrous Spatial Pyramid Pooling)

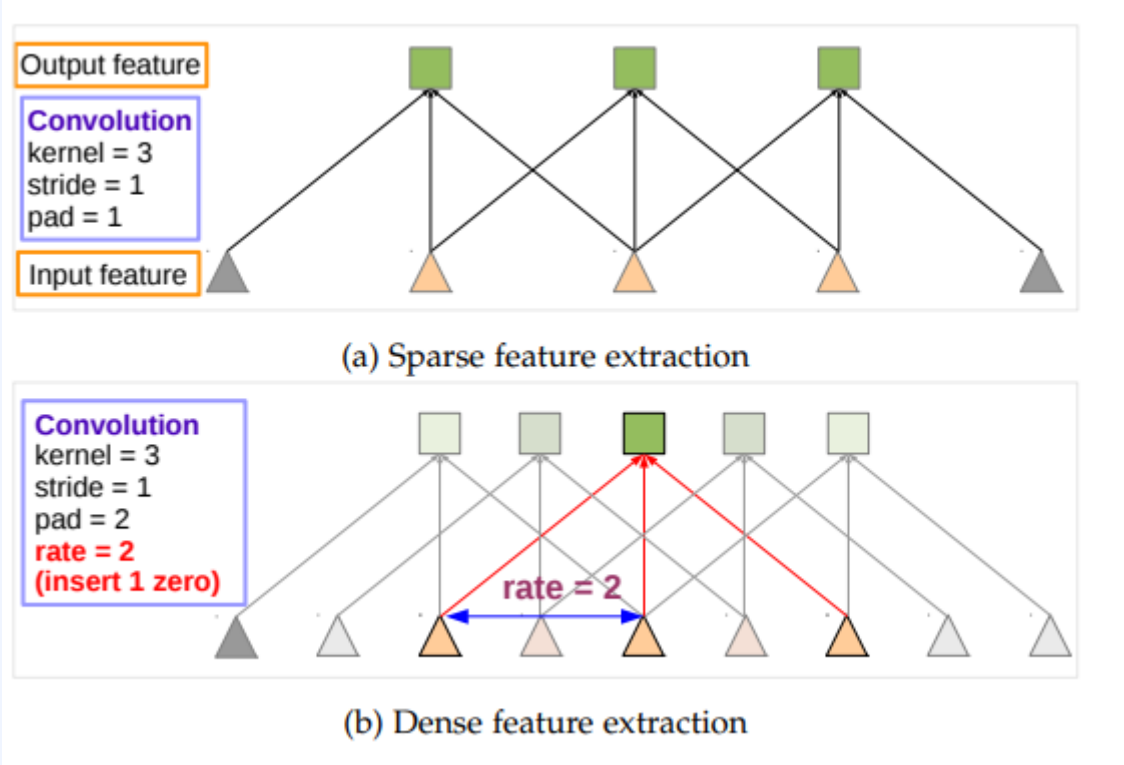


- PSPNet(Pyramid Scene Parsing Network)



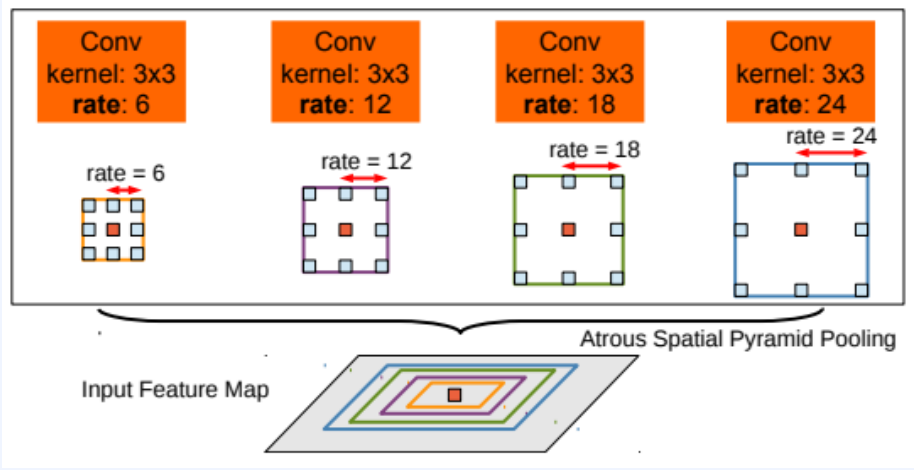
ASPP(Atrous Spatial Pyramid Pooling)

- **Atrous Convolution**

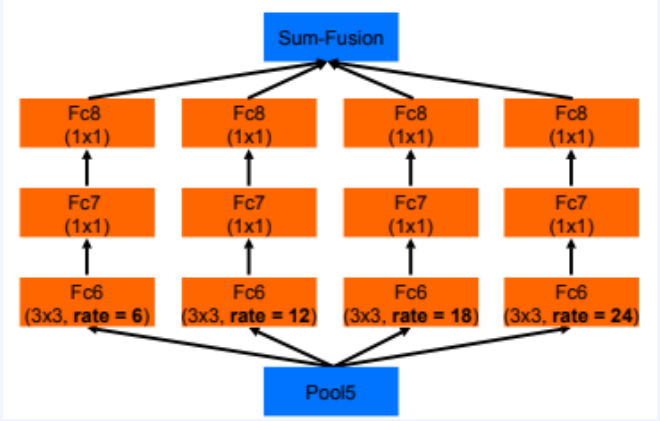


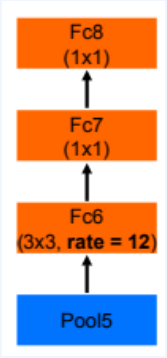
ASPP(Atrous Spatial Pyramid Pooling)

- **Atrous Spatial Pyramid Pooling**



- **DeepLab-ASSP**



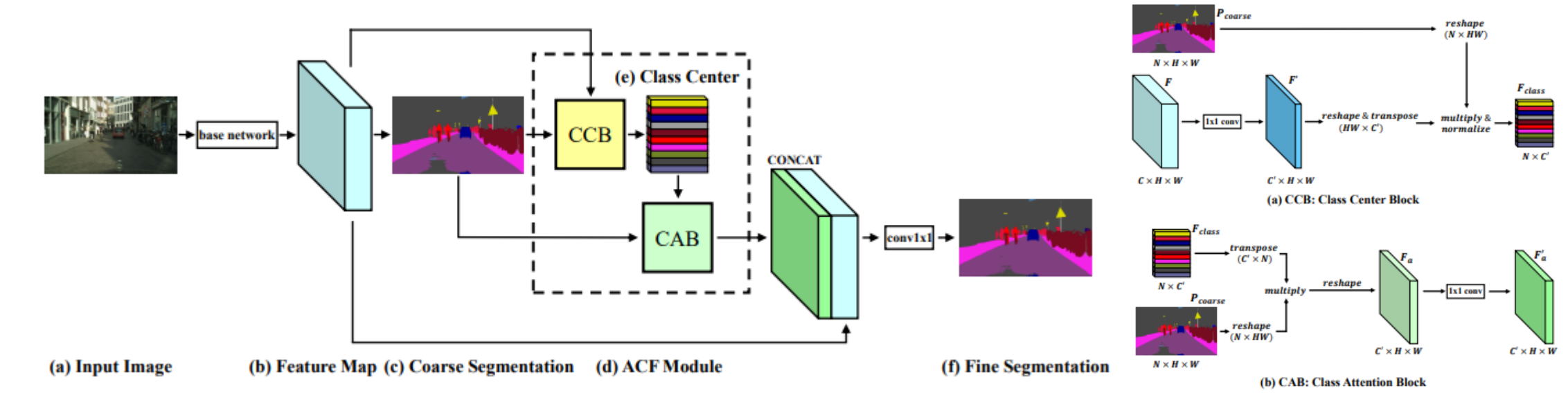


DeepLap

Relation context

Relation context

- Relational context
  - DANet, CFNet, OCNet
    - Self-attention을 통해 픽셀간의 관계를 고려
  - ACFNet
    - Pixel을 Region Set으로 그룹화

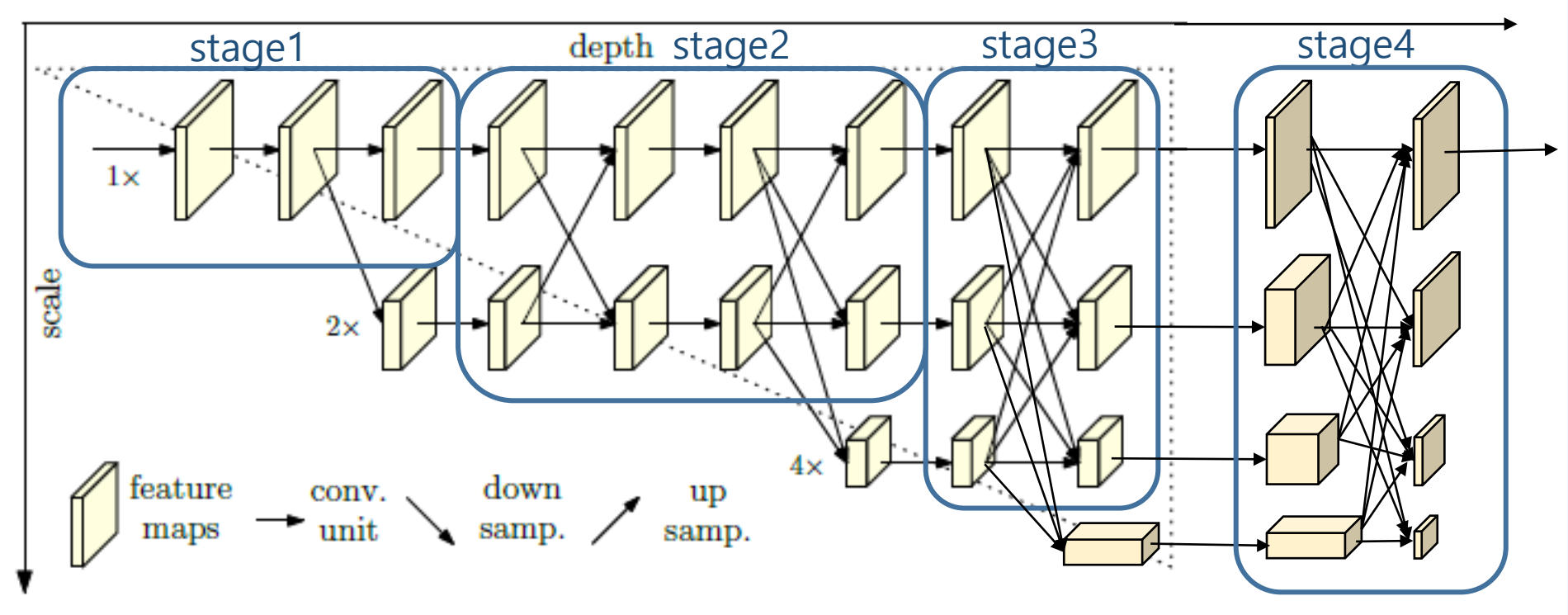


HRNet v2 + OCR

HRNet v2 + OCR

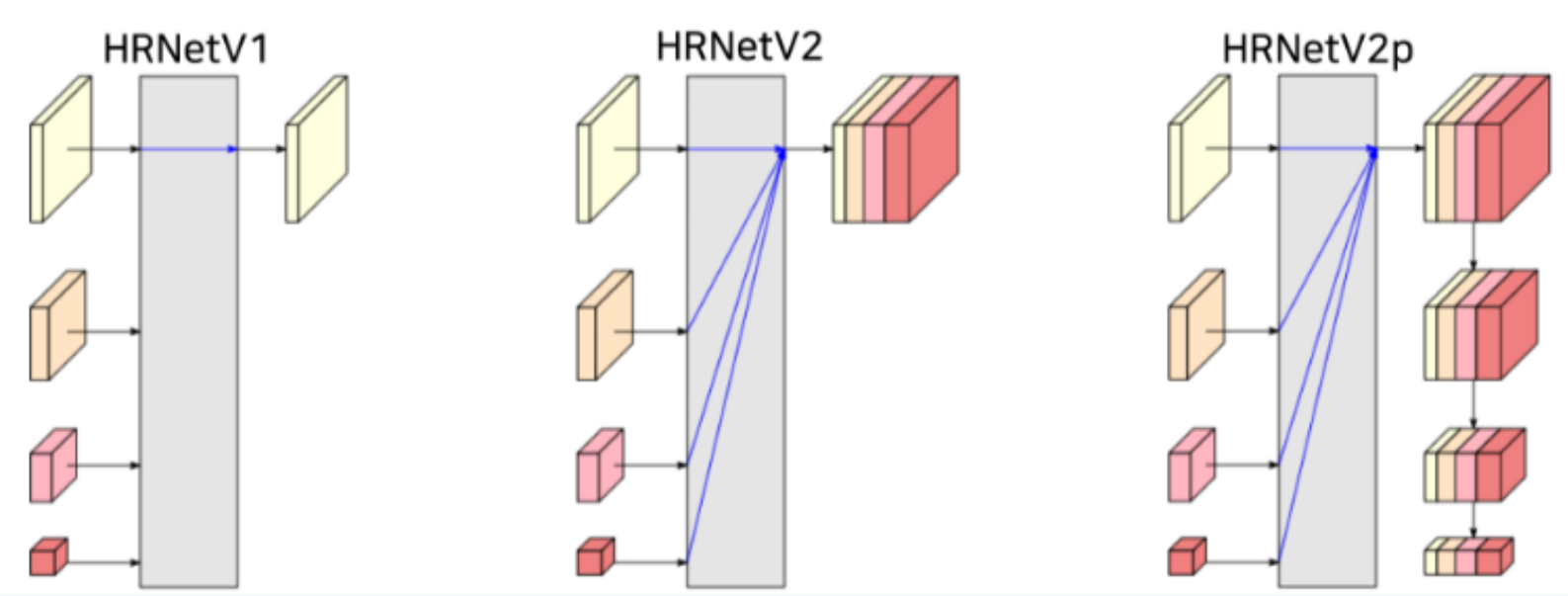
- Semantic segmentation
  - Pixel이 속하는 object의 class를 예측하는 문제(pixel-level dense prediction)
  - HRNet v2 + OCR(proposed) + SegFix -> 1st place on the Cityscapes leaderboard
- OCR-module –three steps
  - Ground truth로 부터 soft object region을 학습
  - Object region에 있는 pixel들을 aggregate함으로써 object region을 계산
  - 각각의 pixel과 object region의 관계를 계산하여 object-contextual representation을 계산하여 pixel representation을 augment
- Segmentation Transformer
  - OCR approach를 transformer encoder-decoder framework로 해석한 부분을 추가함

HRNet v2(backbone)



HRNet v2(backbone)

- HRNet version별 차이

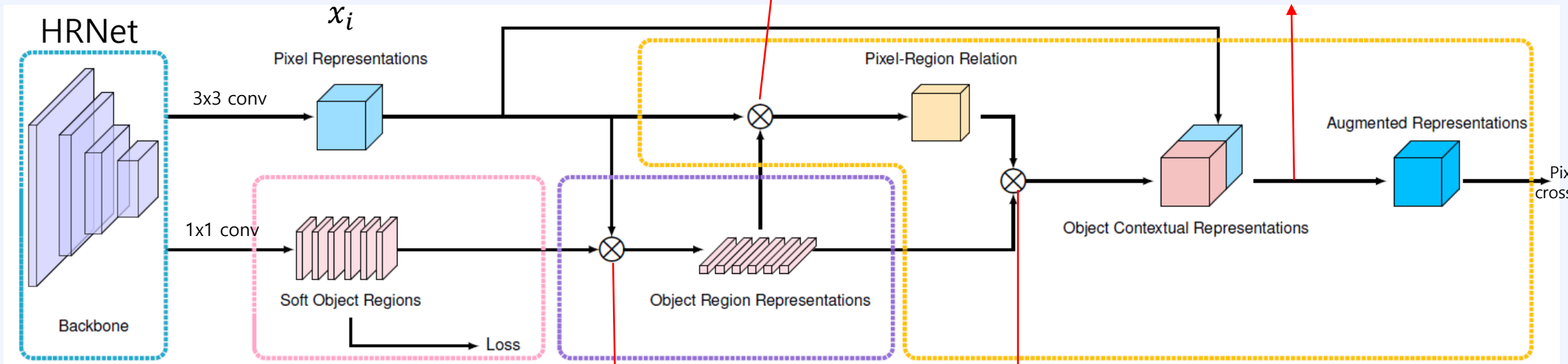


Pipeline of OCR

$$w_{ik} = \frac{e^{k(x_i, f_k)}}{\sum_{j=1}^{K} e^{k(x_i, f_j)}} \cdot k(x, f) = \emptyset(x)^T \varphi(f)$$

$$z_i = g(\left[x_i^T, y_i^T\right]^T)$$

Augmented representations



Pixel-wise cross-entropy loss
Corse segmentation map
$\{M_1, M_2, \ldots, M_k\}$

Degree: corresponding pixel belongs to the class k

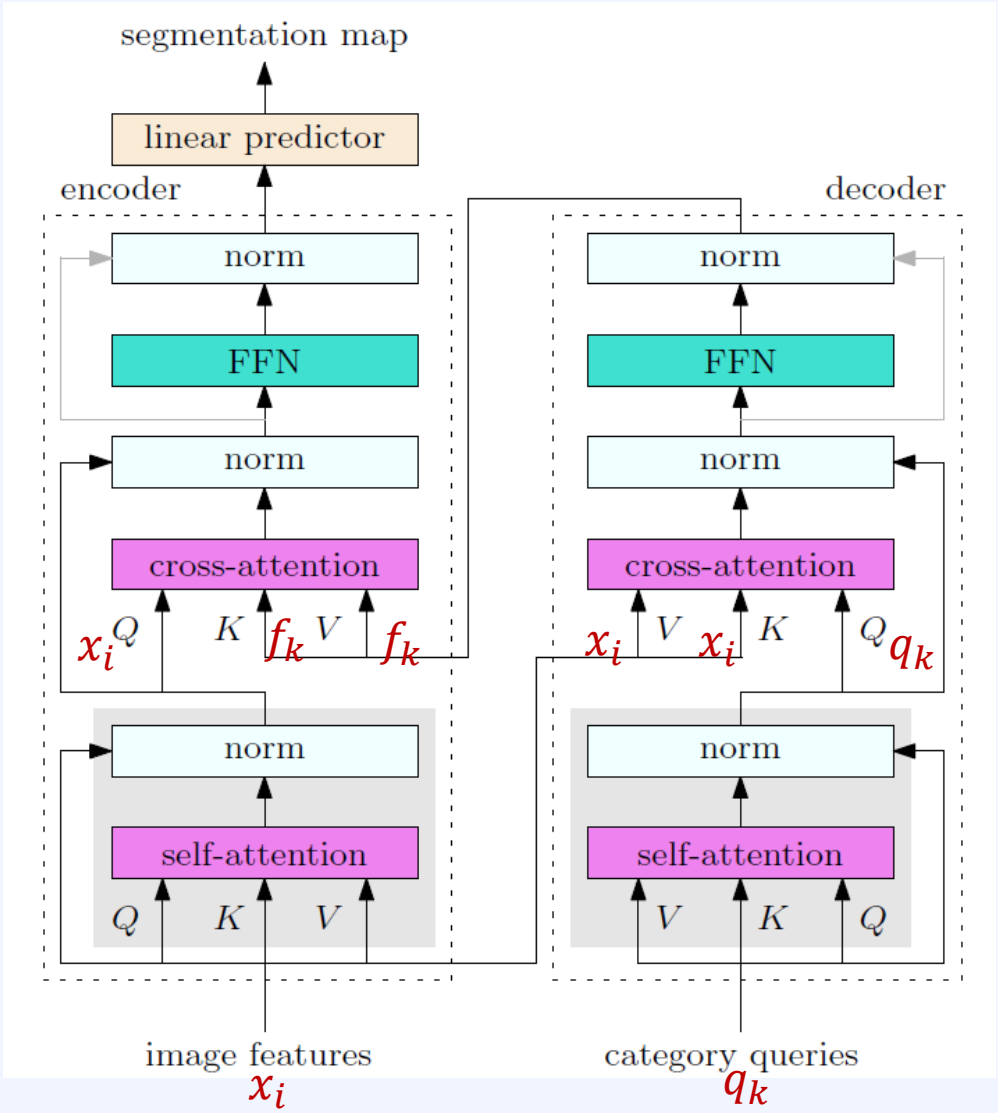
$$f_k = \sum_{i \in \tau} \tilde{m}_{ki} x_i$$

$$y_i = p(\sum_{k=1}^{k} w_{ik} \delta(f_k))$$

Segmentation transformer

$$y_i = p(\sum_{k=1}^{k} w_{ik}\delta(f_k))$$

$$w_{ik} = \frac{e^{k(x_i, f_k)}}{\sum_{j=1}^{K} e^{k(x_i, f_j)}}$$



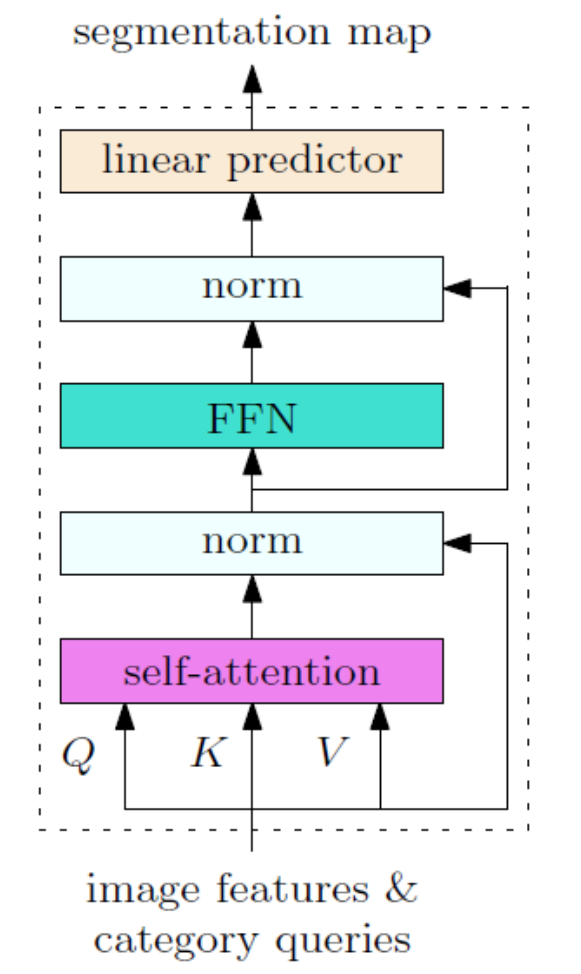
$$f_k = \sum_{i \in \tau} \tilde{m}_{ki} x_i$$

$q_k \rightarrow M_k$(soft object regions)

↓

$\tilde{m}_{ki}$: spatially softmax-normalized

Alternative of segmentation transformer



Summary

- Semantic segmentation
- Multi-scale context
- Relation context
- HRNet+OCR