DeepIntoDeep

CV 집중 탐색 - Image Segmentation

발표자: 김승현

CV 집중 탐색 - Image Segmentation

김승현

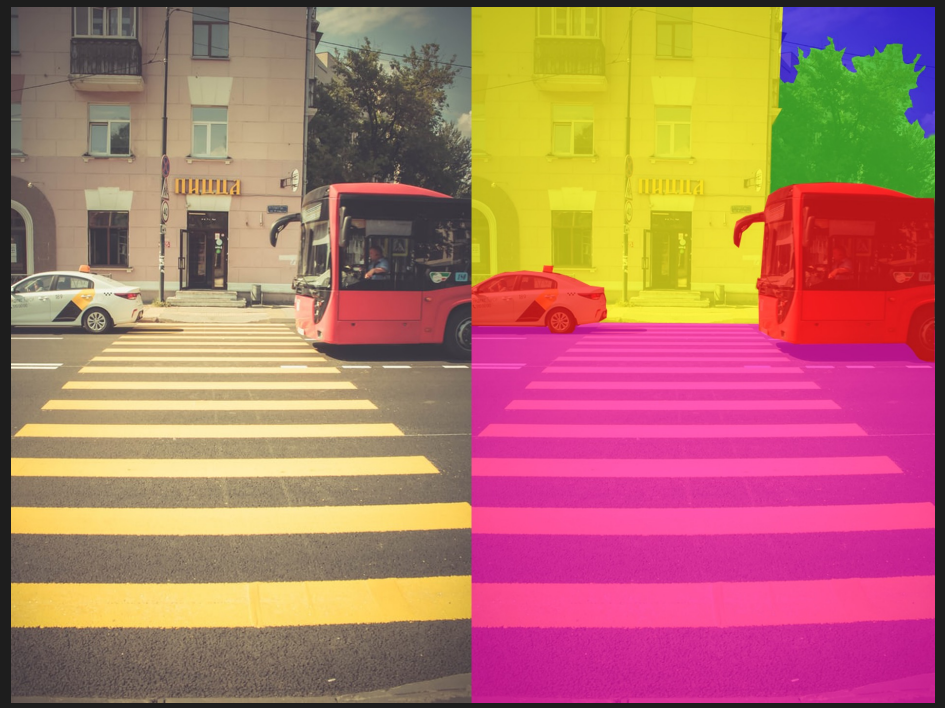
Artificial Intelligence in Korea University(AIKU)

Department of Computer Science and Engineering, Korea University

Image Segmentation

Image Segmentation

- To divide an image into meaningful and distinguishable regions or objects
- An essential component in many visual understanding systems
  - The segmented image is more meaningful and easier to analyze



0 - background

1 - vehicle

2 - road

3 - building

4 - tree

Image Segmentation: Types

Semantic Segmentation

- To classify each pixel with semantic labels

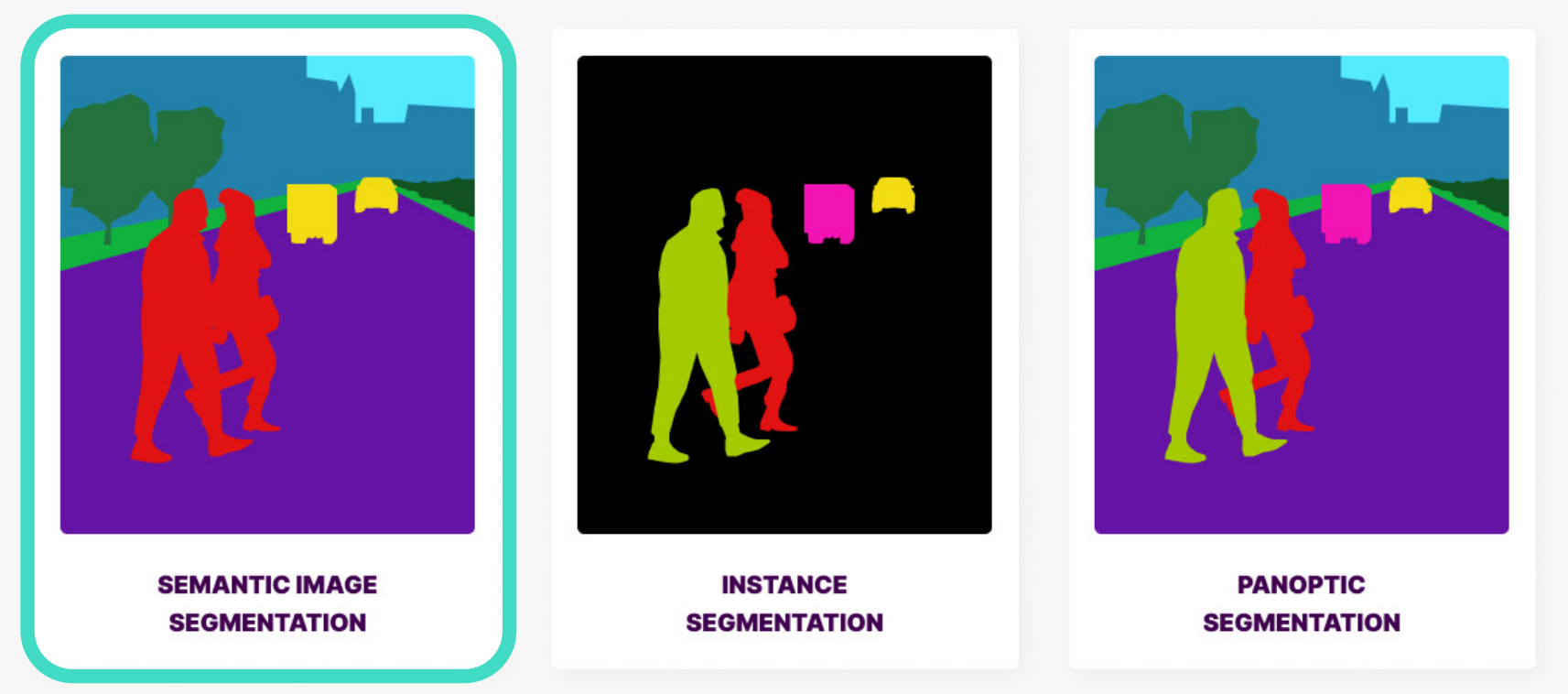


Image Segmentation: Types

Instance Segmentation

- To detect and segment individual objects

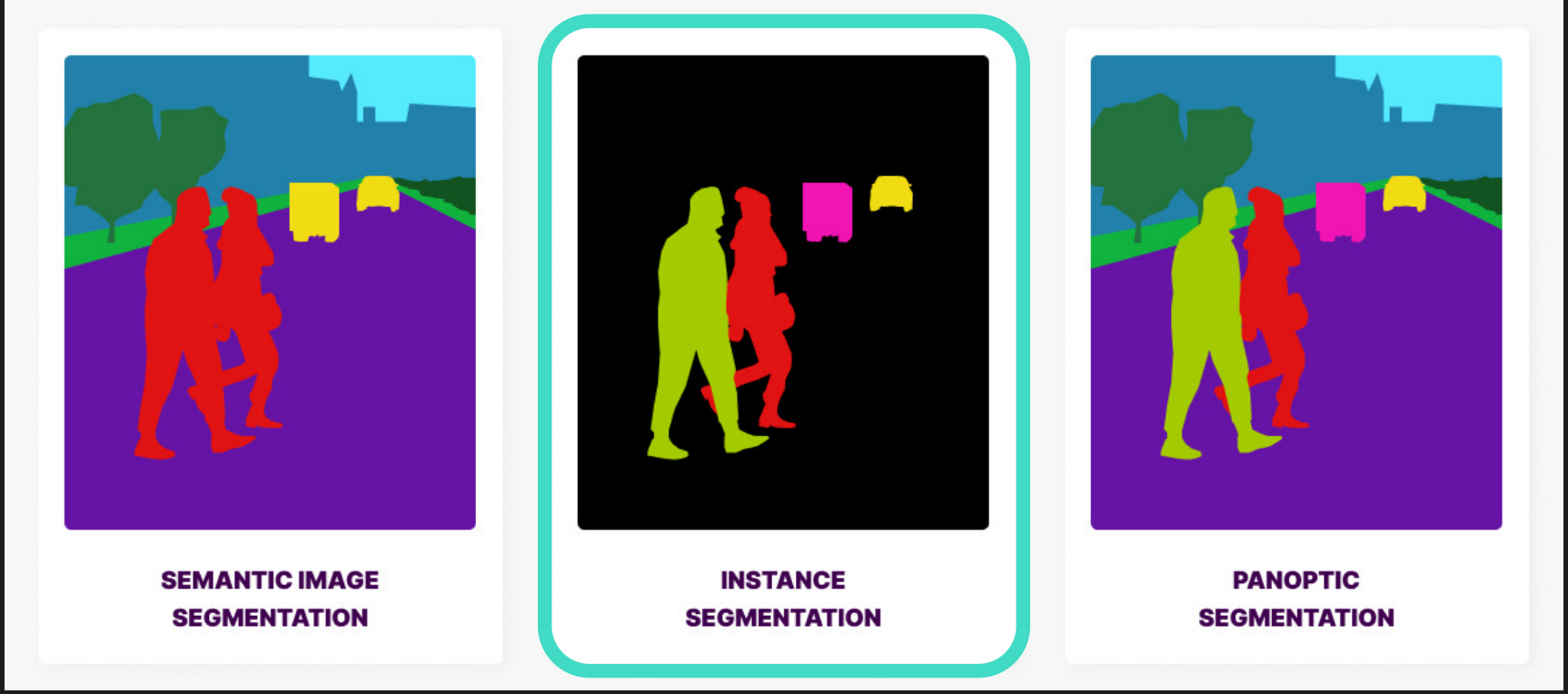


Image Segmentation: Types

Panoptic Segmentation

- Semantic segmentation + Instance segmentation

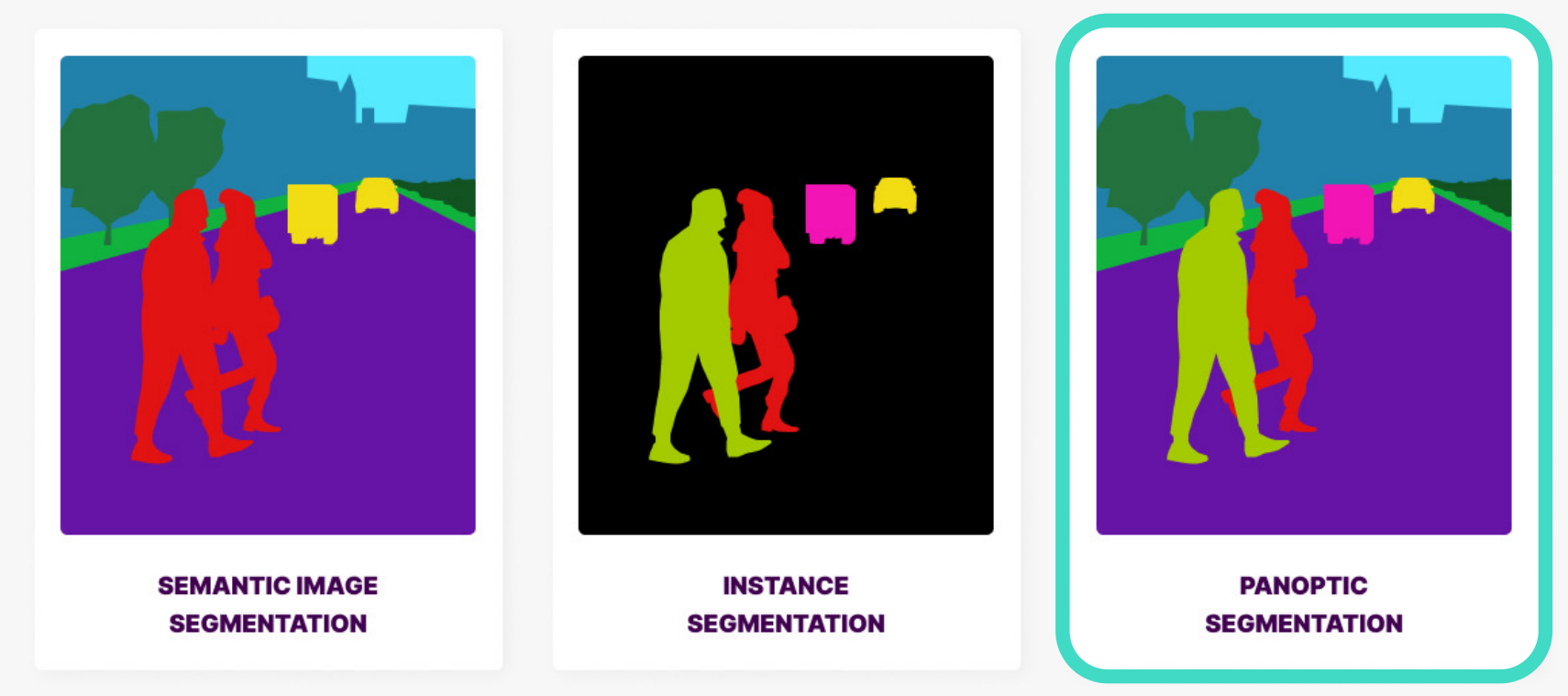


Image Segmentation: How-to

Traditional Methods

- Thresholding, histogram-based bundling, region growing, k-means clustering, and watersheds
- More advanced algorithms: Active contours, graph cuts, conditional and Markov random fields, and sparsity-based methods

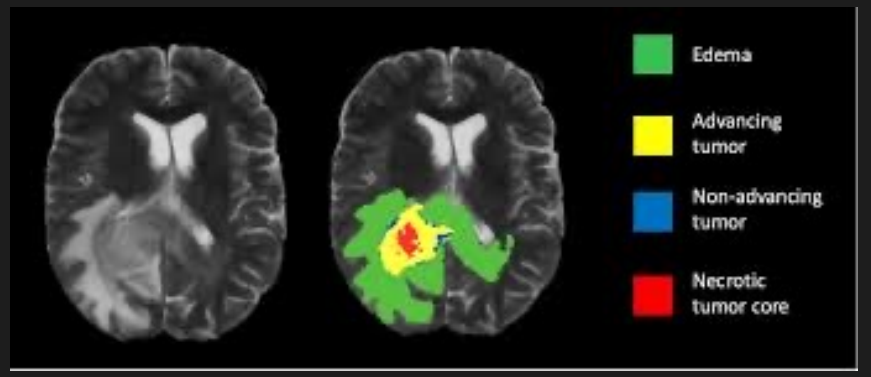
Deep Learning-based Methods

- Utilize a pretrained image classification model as an encoder

Image Segmentation: Applications

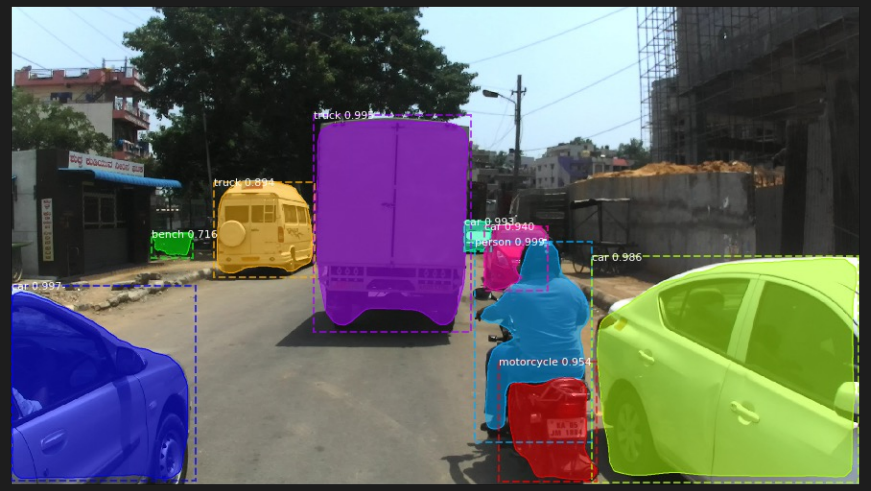
Medical image analysis

- Tumor boundary extraction
- Tissue volume measurement



Autonomous vehicles

- Navigable surface and pedestrian detection



Video surveilance

Augmented reality

Remote sensing

- Satellite image segmentation

Image Segmentation: Applications

Portrait mode on Google Pixel smartphone

- Separate the image into foreground (typically a person) and background layers
- Then blur the background layer

https://ai.googleblog.com/2017/10/portrait-mode-on-pixel-2-and-pixel-2-xl.html

Image Segmentation: Datasets

2D Datasets

- PASCAL Visual Object Classes (VOC)
- PASCAL Context
- Microsoft Common Objects in Context (MS COCO)
- Cityscapes
- KITTI

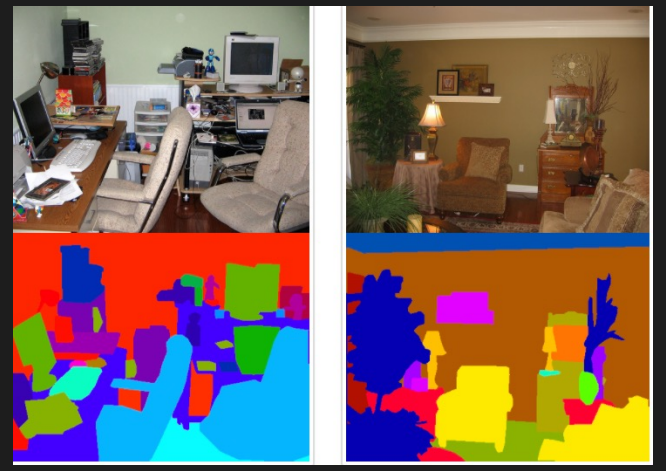
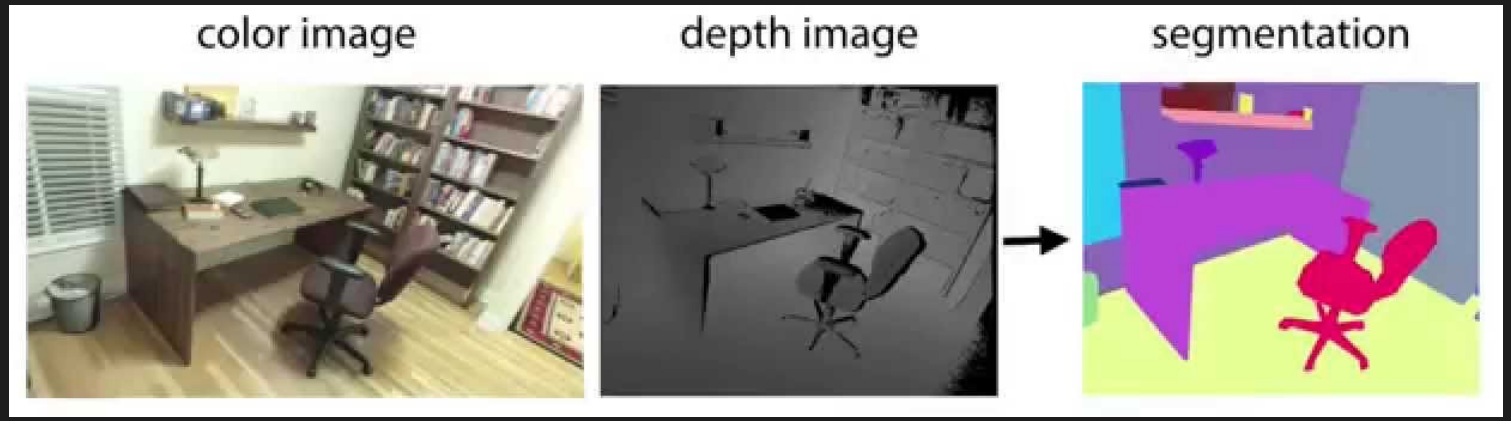


Image Segmentation: Datasets

2.5D Datasets (RGB-D)

- NYU-D V2
- SUN 3D
- SUN RGB-D



3D Datasets

- Stanford 2D-3D
- ShapeNet Core
- Sydney Urbun Objects Dataset

Image Segmentation: Evaluation

Pixel Accuracy

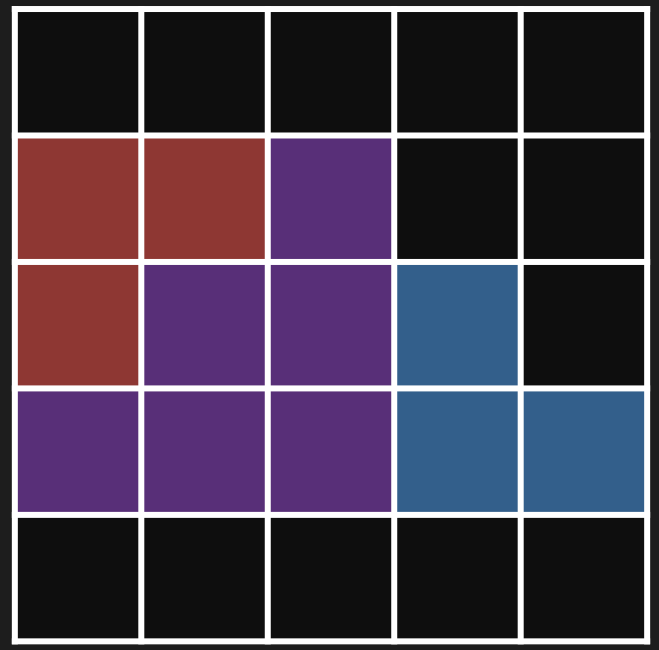
- The number of correctly-classified pixels, divided by the total number of pixels

$$PA = \frac{Correct}{Total} = \frac{6}{25}$$

Image Segmentation: Evaluation

Intersection over Union (IoU) (Jaccard Index)

- One of the most commonly-used metrics in semantic segmentation
- The intersection area of the predicted segmentation and the ground truth, divided by their union area



$$\mathrm{IOU} = \frac{Intersect}{Union} = \frac{6}{12}$$

Fully Convolutional Network

Naïve Solution Idea

- We should output a “dense” segmentation map
- Image classification just predict labels; How can we solve this task?

Input:
3 x H x W

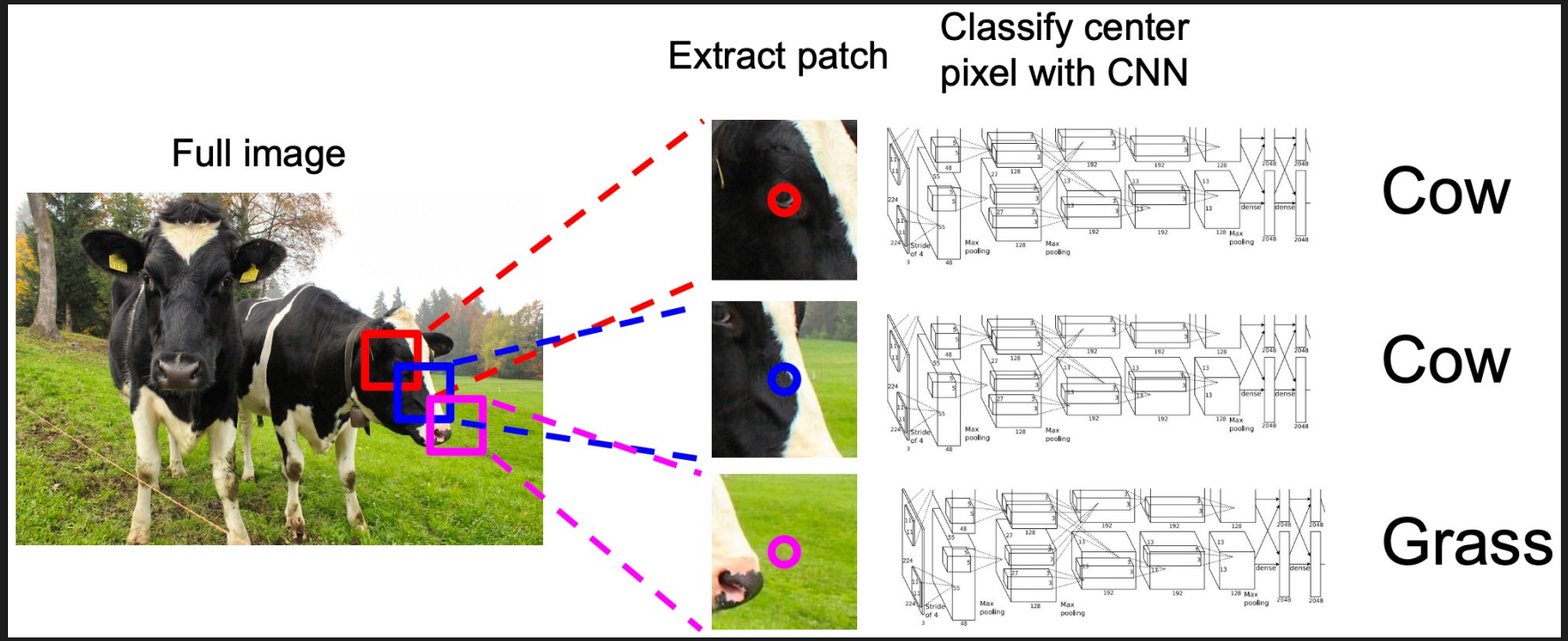
Predictions:
H x W

Naïve Solution Idea

1. Classify each pixel with the corresponding pixel
    - Impossible; We need context information!

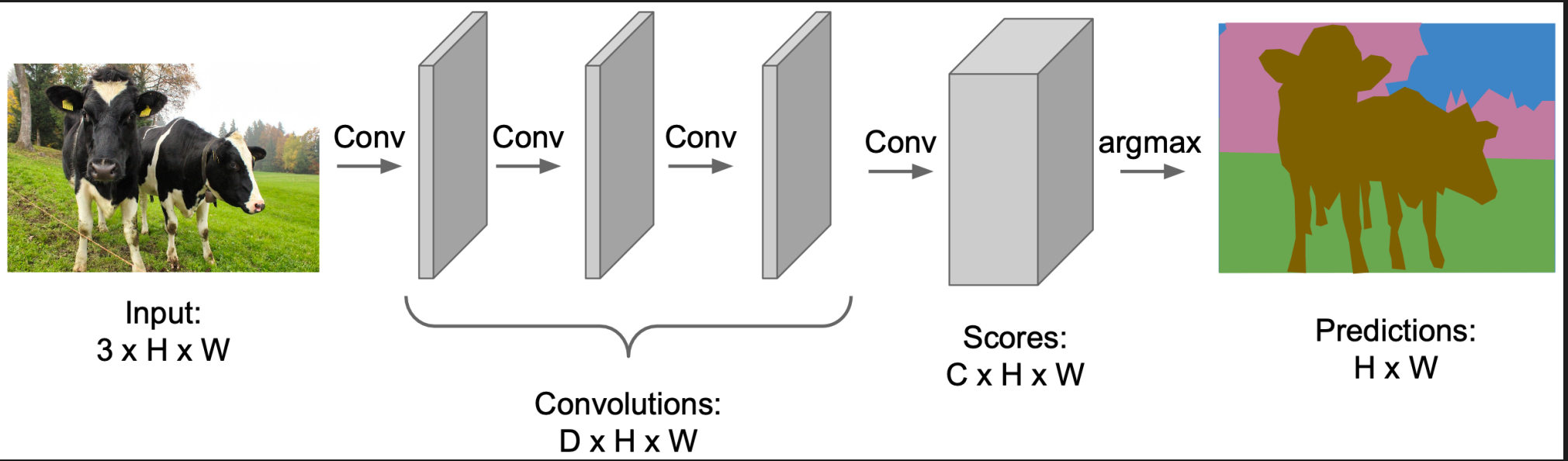
Naïve Solution Idea

2. Classify a center pixel for each patch
   - Inefficient; There are shared regions among patches!



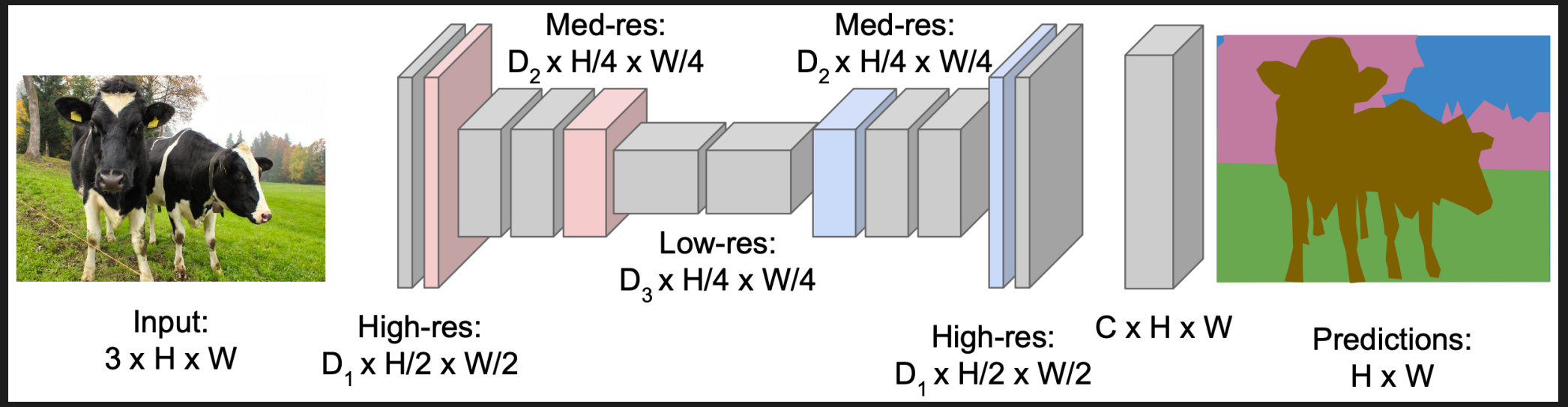
Naïve Solution Idea

3. Classify the whole image without downsampling
   - Operations on the original image resolution will be very expensive…



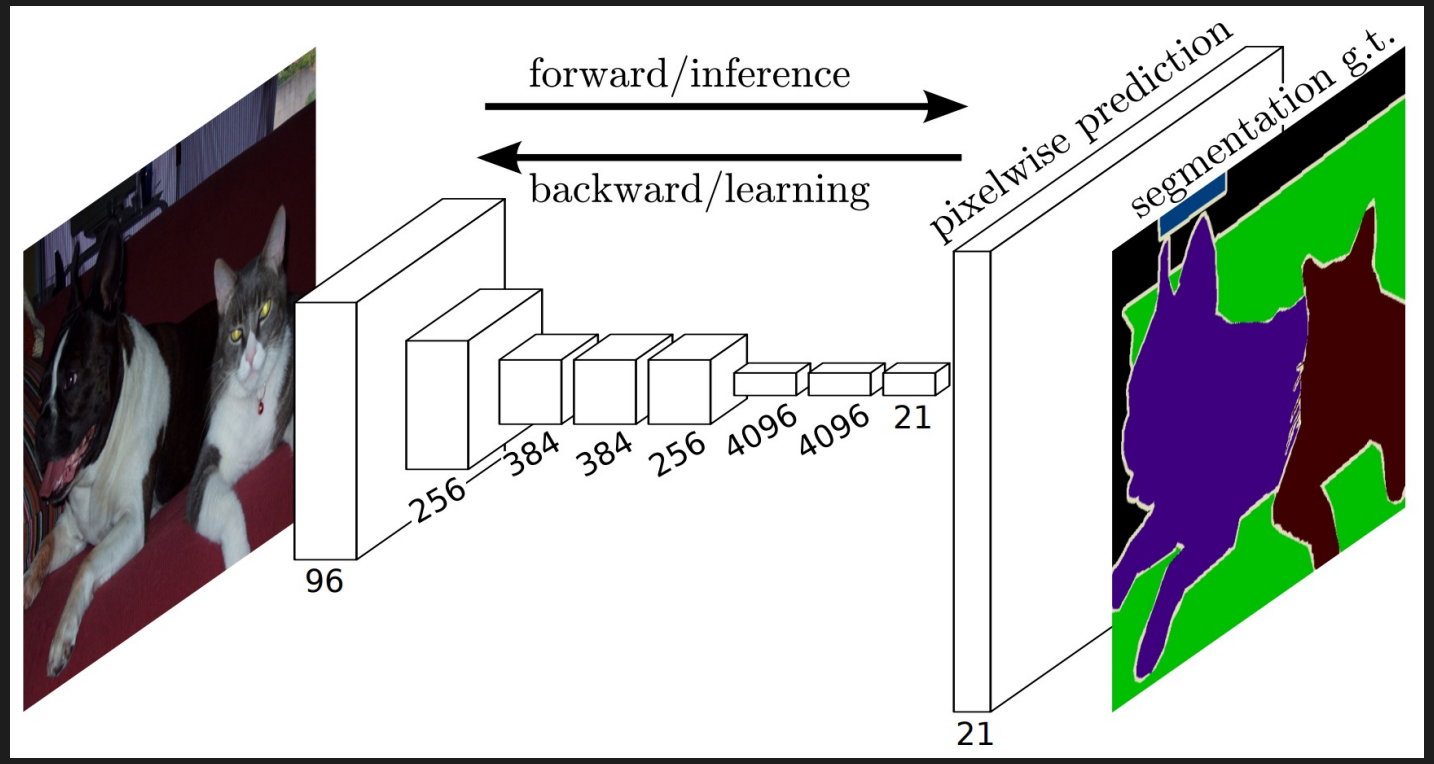
Naïve Solution Idea

4. Classify the whole image with downsampling and upsampling
   - Good!



Fully Convolutional Network

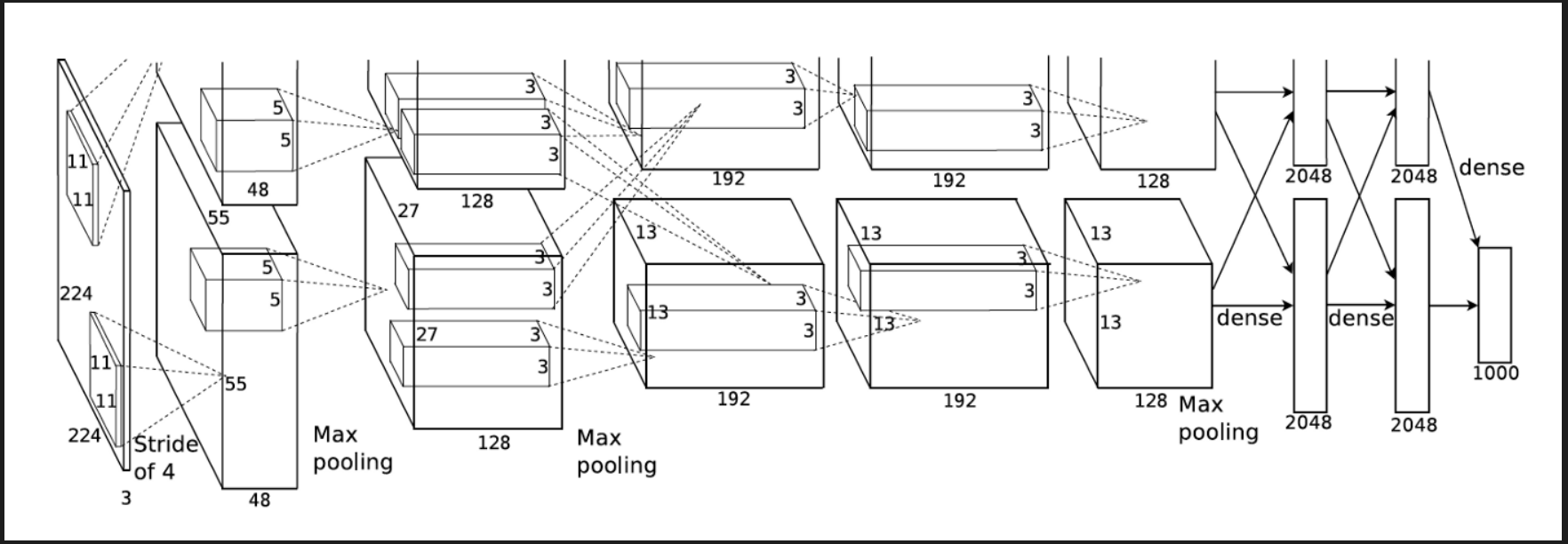
- Composed of convolutional layers; no fully-connected layers
- Upsampling operations to match the resolution



Fully Convolutional Network

Convolutionalization

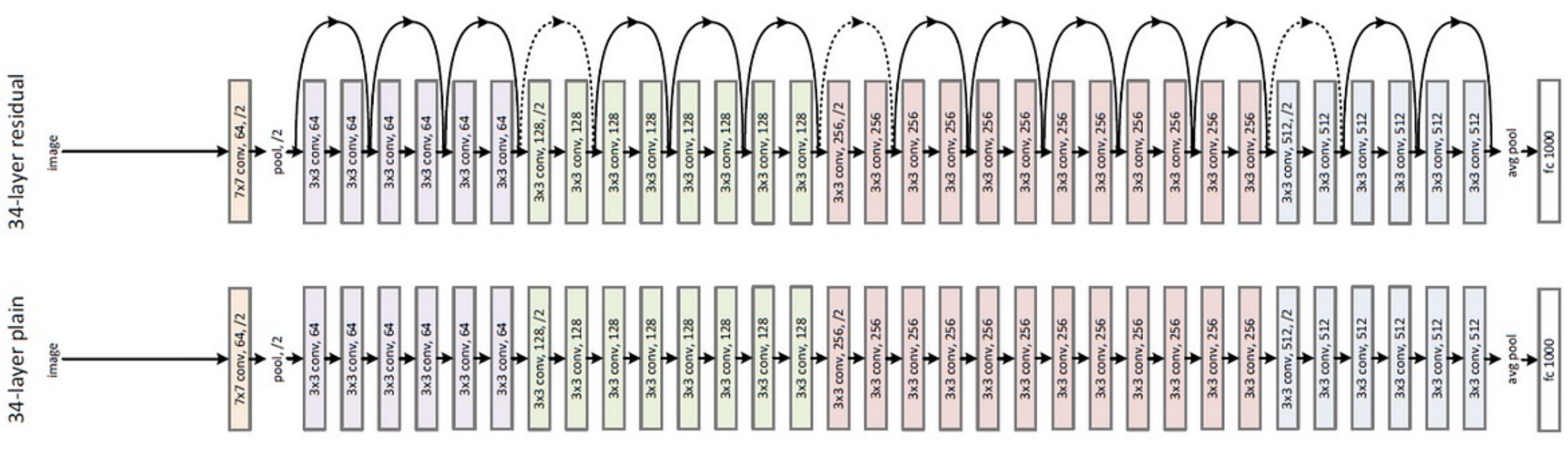
- AlexNet



Fully Convolutional Network

Convolutionalization

- ResNet



Fully Convolutional Network

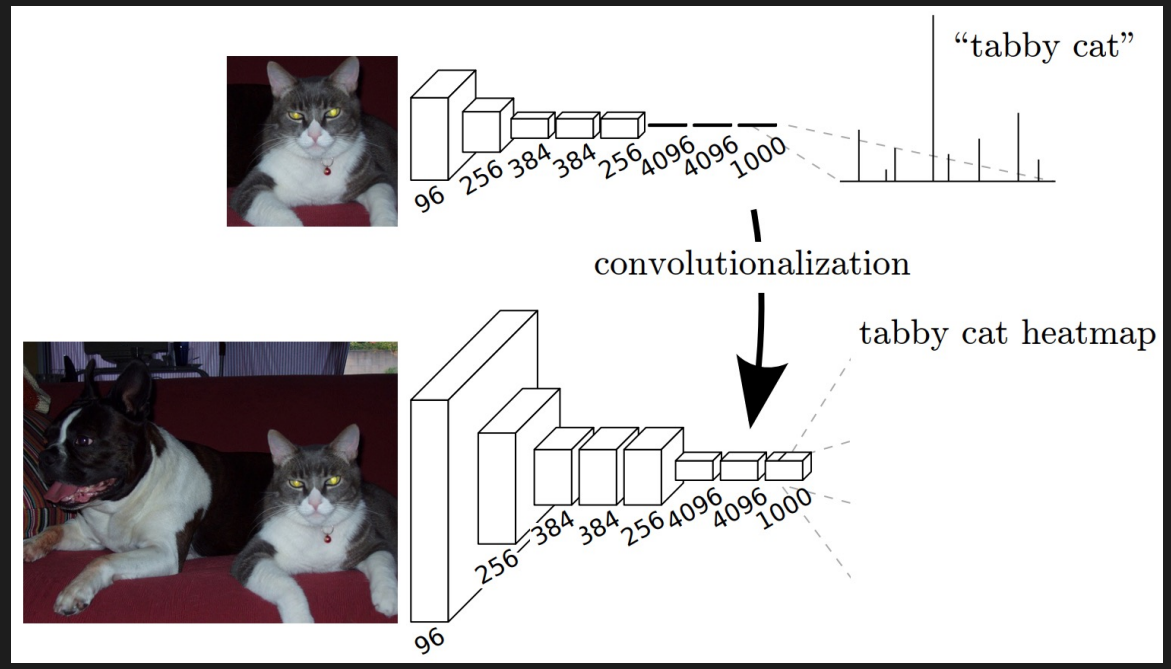
Convolutionalization

- Fully Convolutional Network (FCN) is a modified image classification model
- Fully-connected layers in the original image classification model
  - To classify the extracted features
- Spatial invariance and fixed input size are not useful in image segmentation

Fully Convolutional Network

Convolutionalization

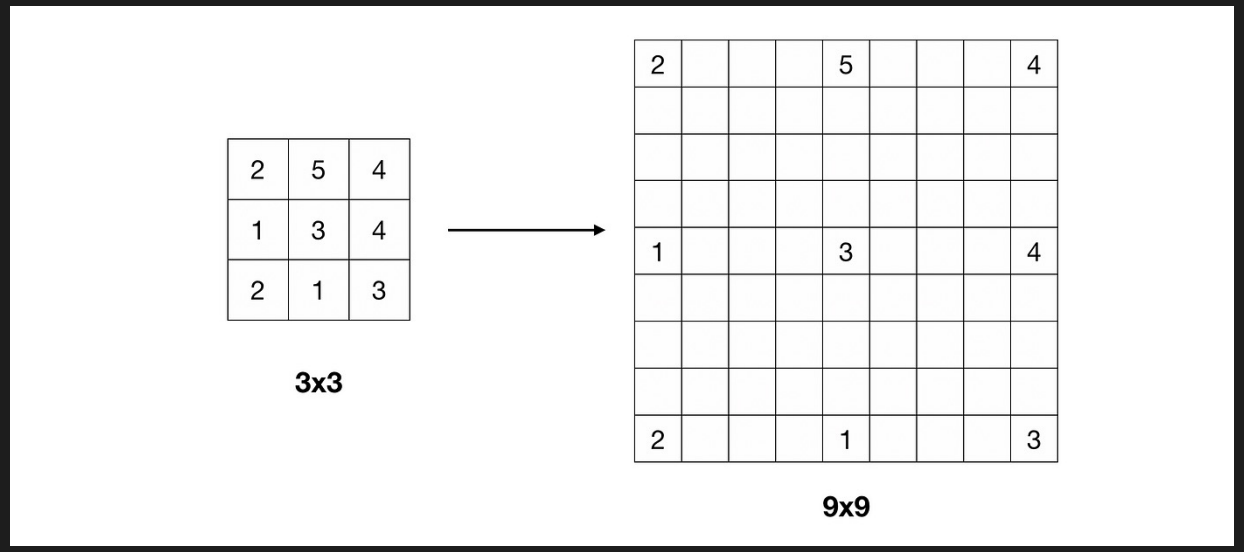
- Replace all the fully-connected layers into convolutional layers
- Spatial information remains



Fully Convolutional Network

Upsampling

- The final feature map is 1/32 of the original image size
- Converting the coarse map to a dense map is necessary
- Upsampling through bilinear interpolation and deconvolution

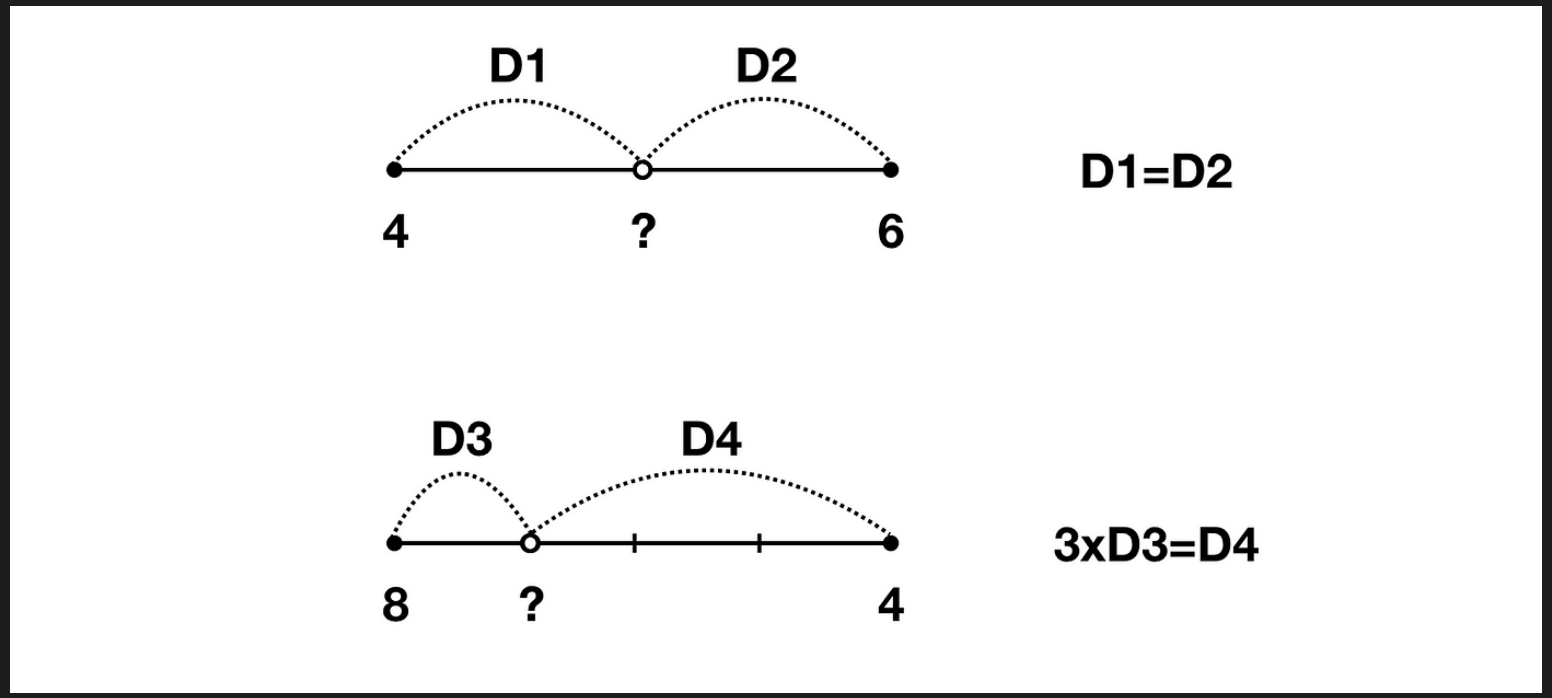


https://medium.com/@msmapark2/fcn-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-fully-convolutional-networks-for-semantic-segmentation-81f016d76204

Fully Convolutional Network

Upsampling

- Linear interpolation

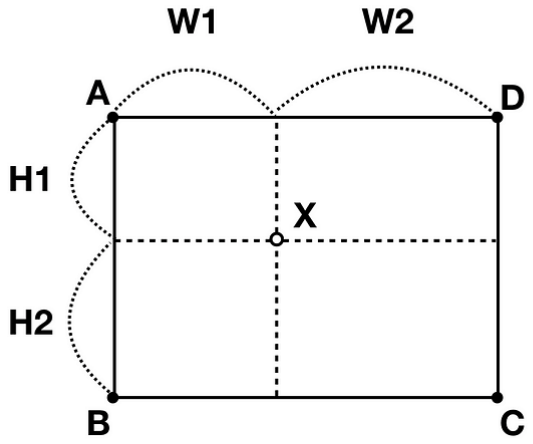


https://medium.com/@msmapark2/fcn-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-fully-convolutional-networks-for-semantic-segmentation-81f016d76204

Fully Convolutional Network

Upsampling

- Bilinear interpolation



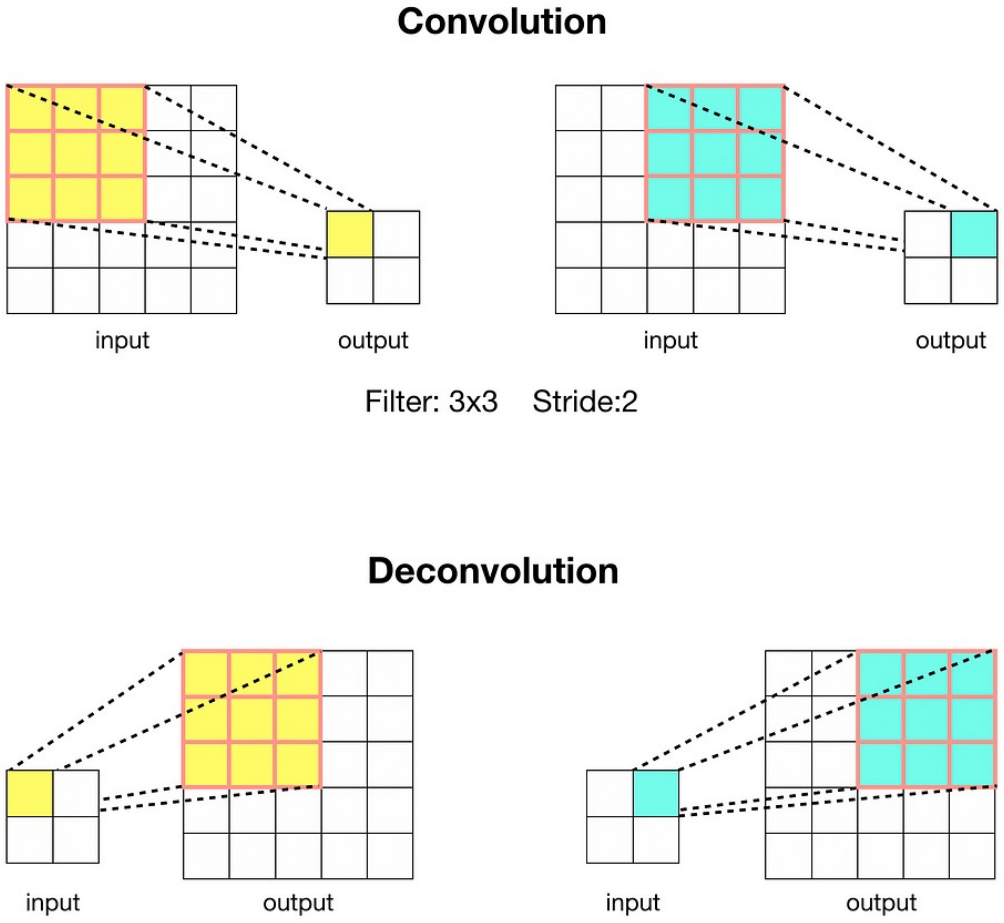
$$X = \left(A\frac{H2}{H1+H2} + B\frac{H1}{H1+H2}\right)\frac{W2}{W1+W2} + \left(D\frac{H2}{H1+H2} + C\frac{H1}{H1+H2}\right)\frac{W1}{W1+W2}$$

https://medium.com/@msmapark2/fcn-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-fully-convolutional-networks-for-semantic-segmentation-81f016d76204

Fully Convolutional Network

Upsampling

- Deconvolution

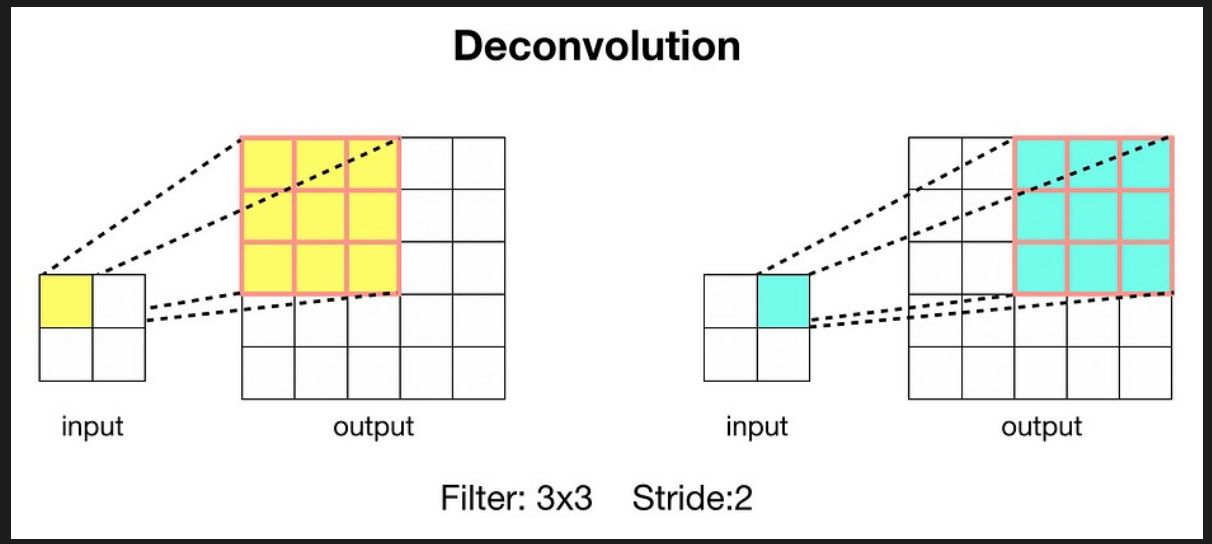


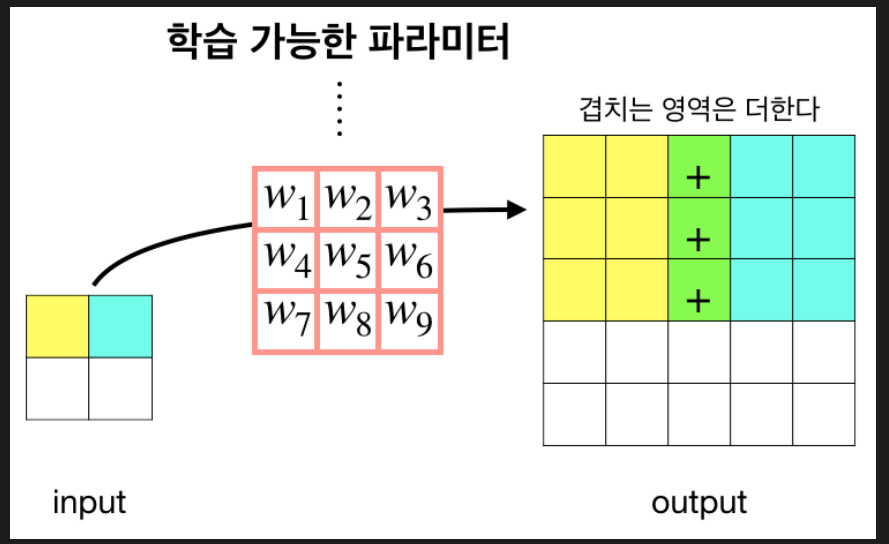
https://medium.com/@msmapark2/fcn-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-fully-convolutional-networks-for-semantic-segmentation-81f016d76204

Fully Convolutional Network

Upsampling

- Deconvolution



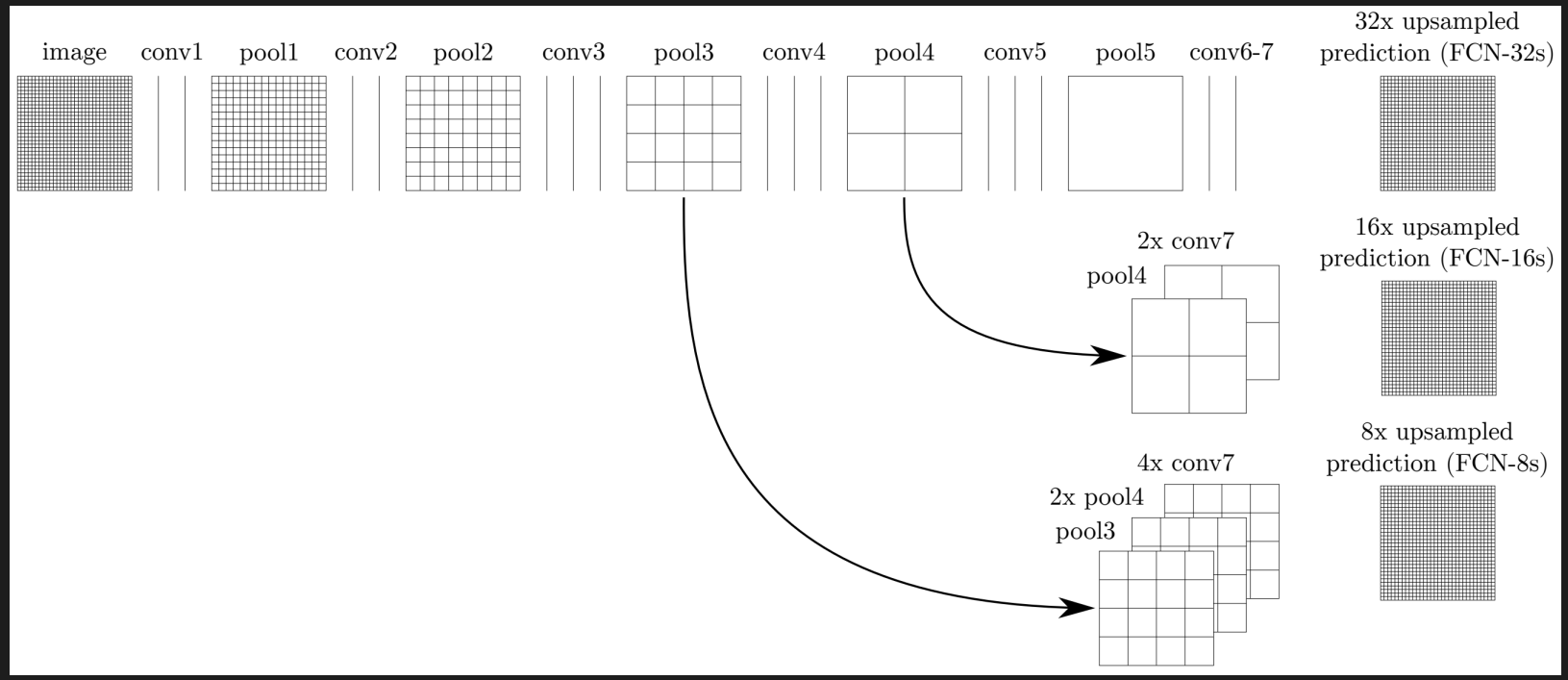


https://medium.com/@msmapark2/fcn-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-fully-convolutional-networks-for-semantic-segmentation-81f016d76204

Fully Convolutional Network

Skip Architecture

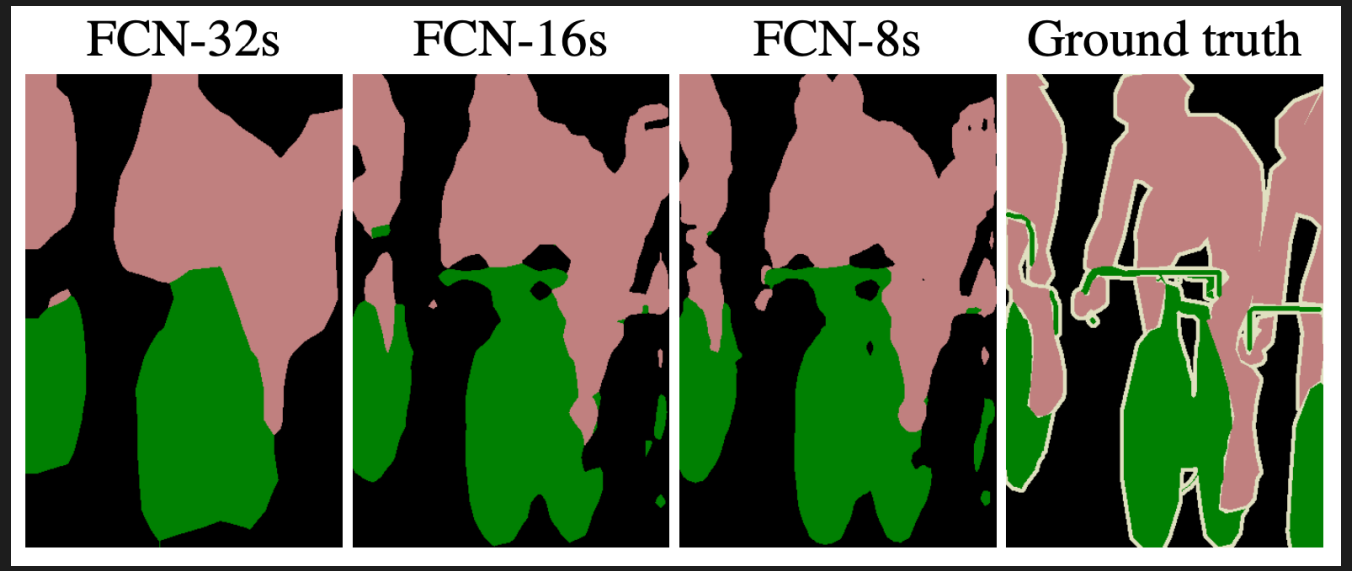
- Feature map from shallow layers: fine, local, location
- Feature map from deep layers: coarse, global, semantic



Fully Convolutional Network

Skip Architecture

- Combine the shallow-level feature map to the deep-level feature map



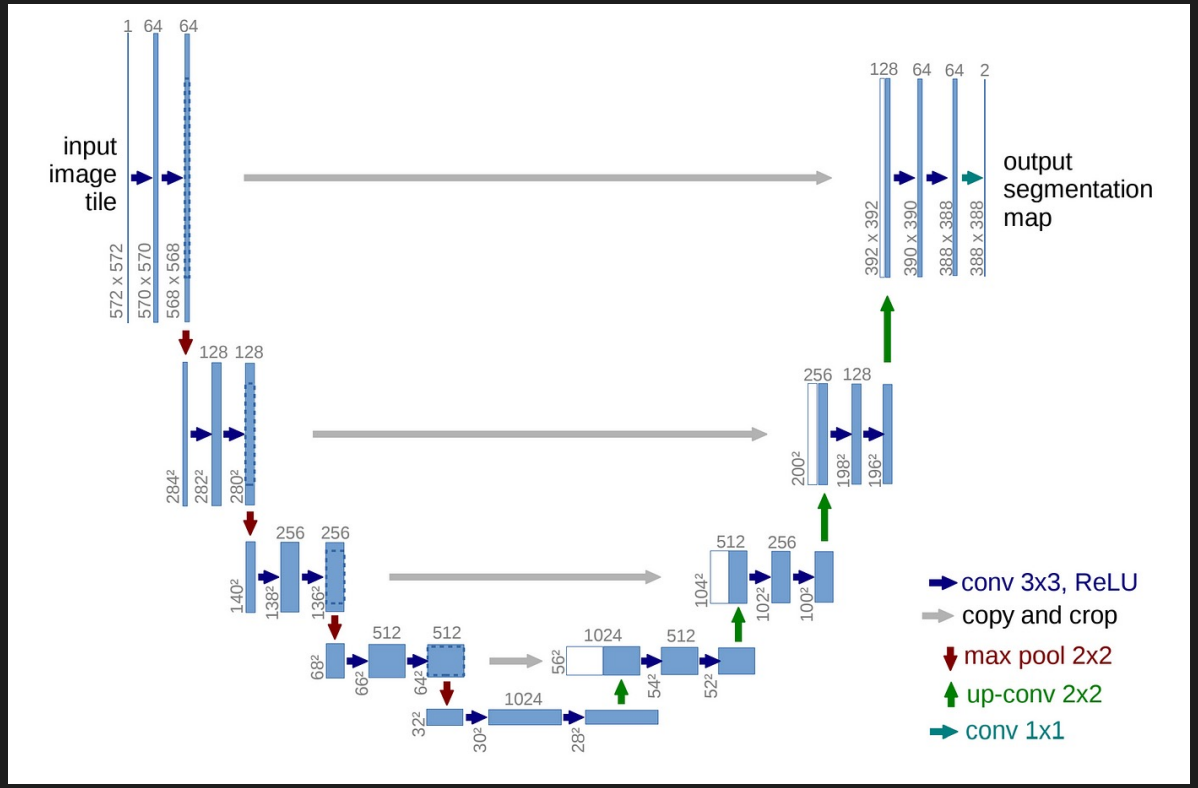
Fully Convolutional Network: Summary

- Modified image classification model
- Fully convolutional for the spatial information
- Upsampling to match the original size
- Skip architecture to utilize both local and global (semantic) information

U-Net

U-Net

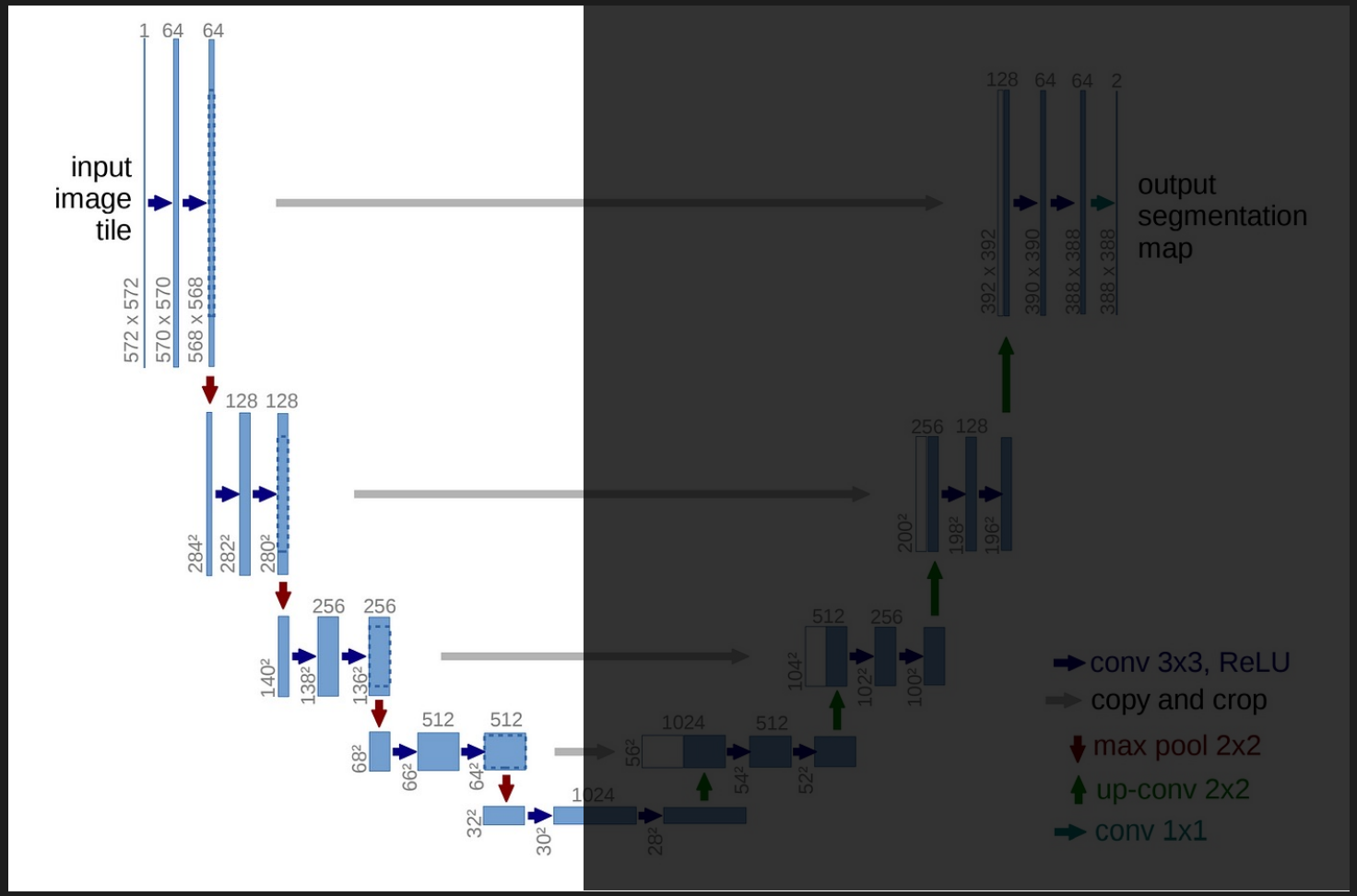
- Suggested for biomedical segmentation tasks
- Symmetric “U” shape → U-Net



U-Net

Contracting Path

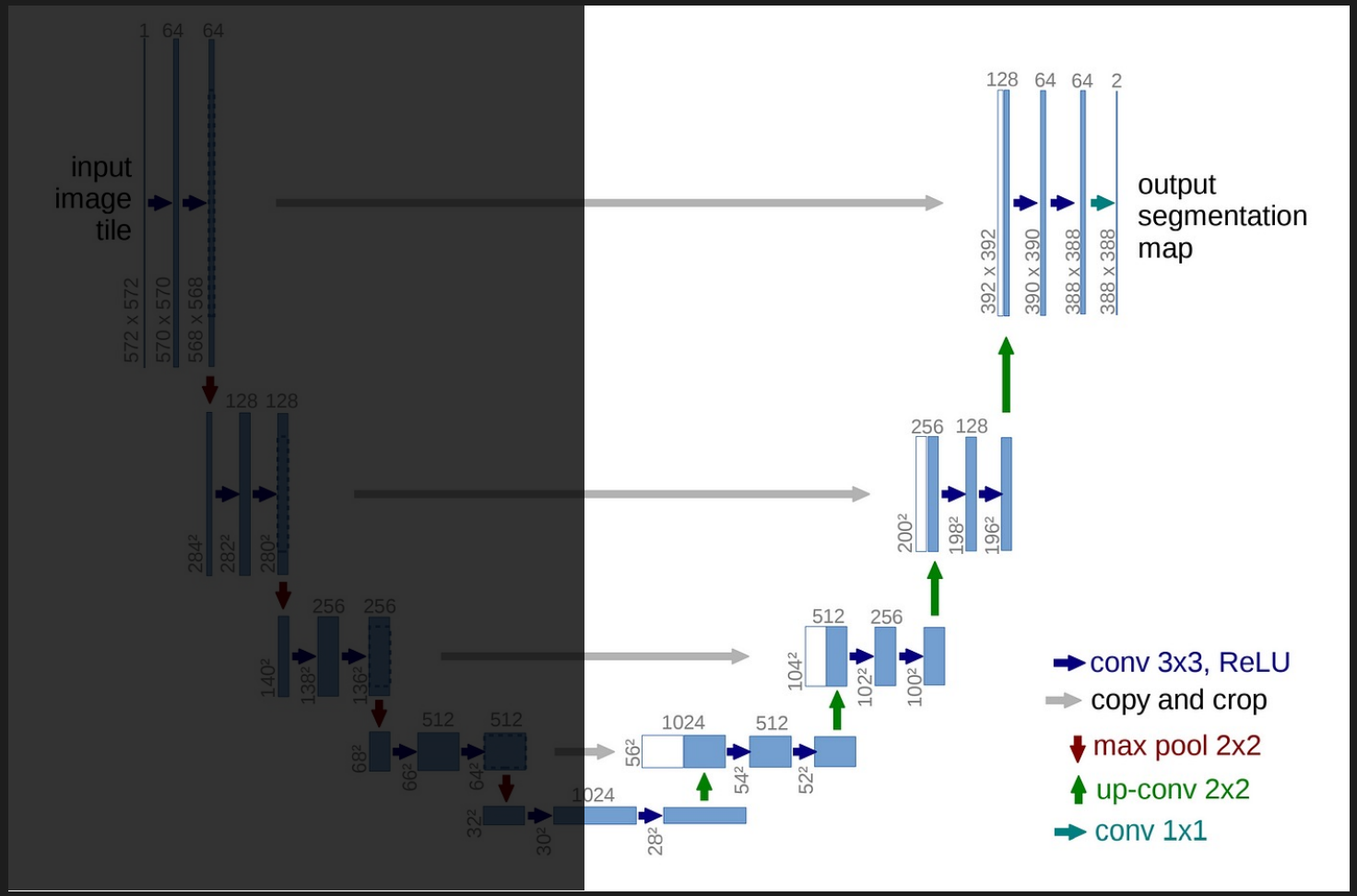
- Reduce the feature map size after two convolution operations
- The number of channels is doubled
- To obtain context features



U-Net

Expanding Path

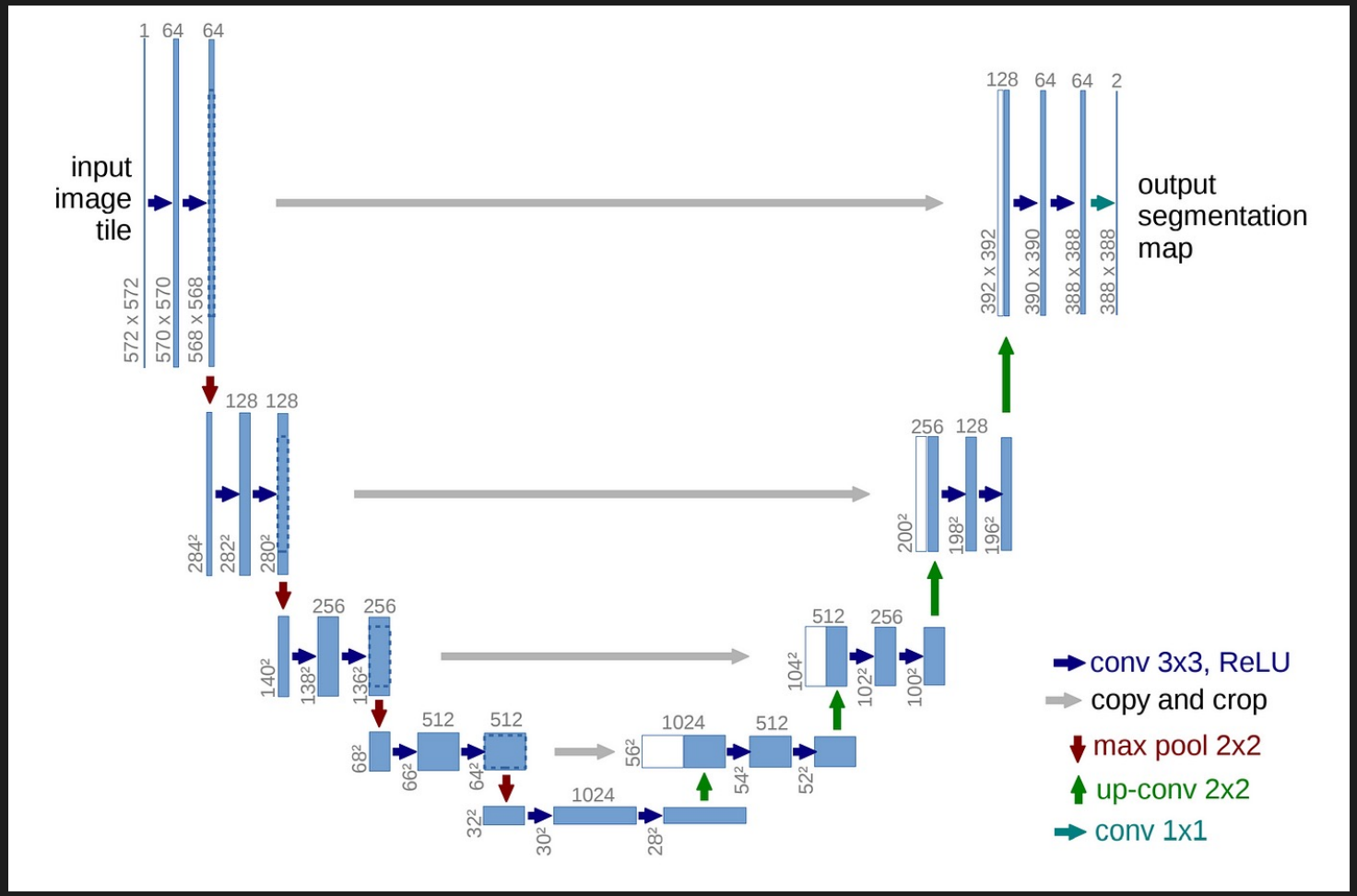
- Increase the feature map size after two convolution operations
- The number of channels is halved
- For localization



U-Net

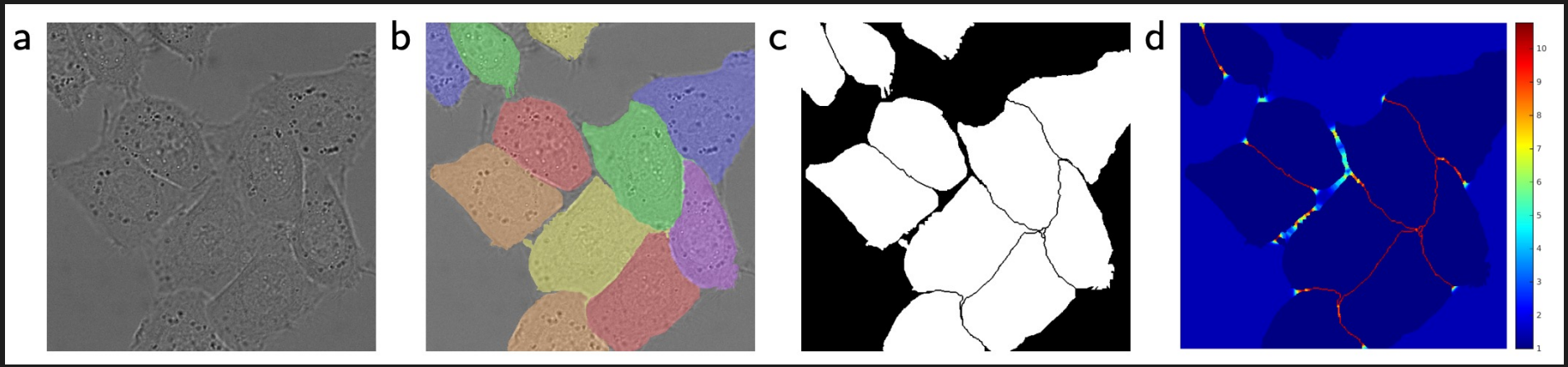
Concatenating Feature Maps

- Concatenate the previous feature map from the coressponding level
- Multiple times
- To obtain rich features



U-Net

- Strong data augmentation
- Won the ISBI cell tracking challenge 2015 only using 30 training images!



U-Net

- Core component of the Stable Diffusion architecture

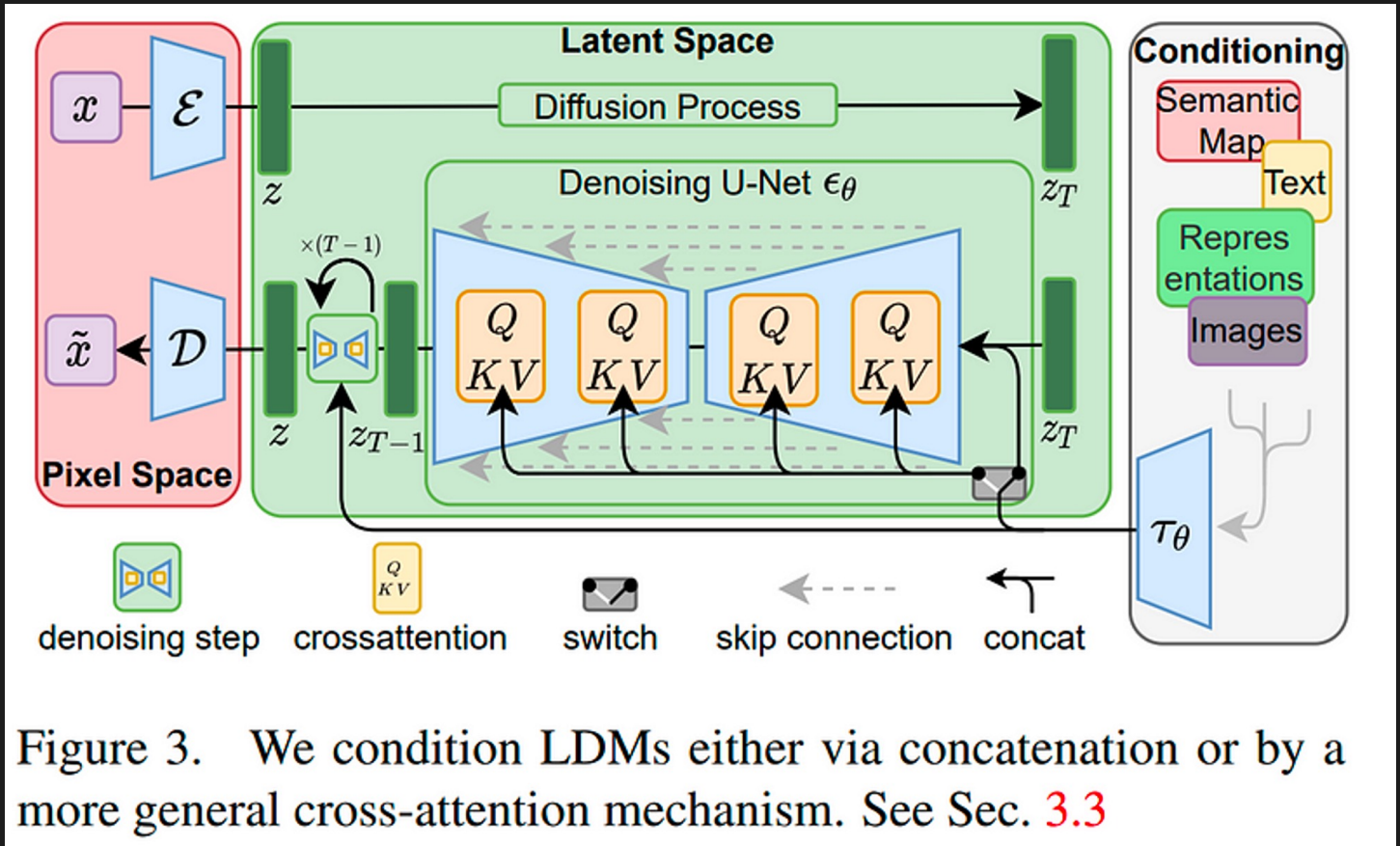


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

U-Net: Summary

- Contracting path and expanding path
- Multiple concatenation with a feature map from the contracting path
- Context and localization information
- Stable Diffusion component

DeepLab Family

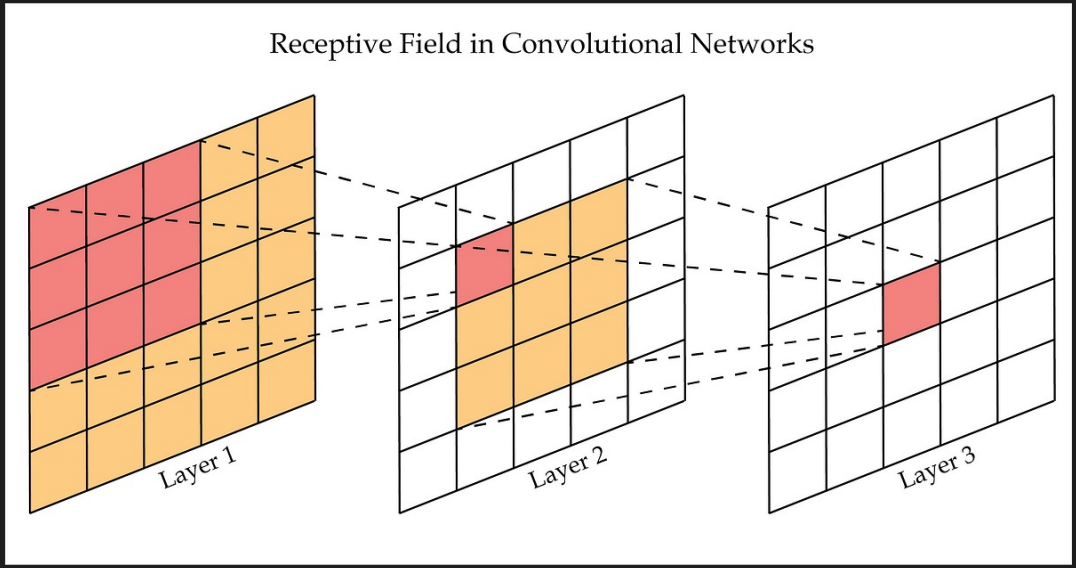
DeepLab Family

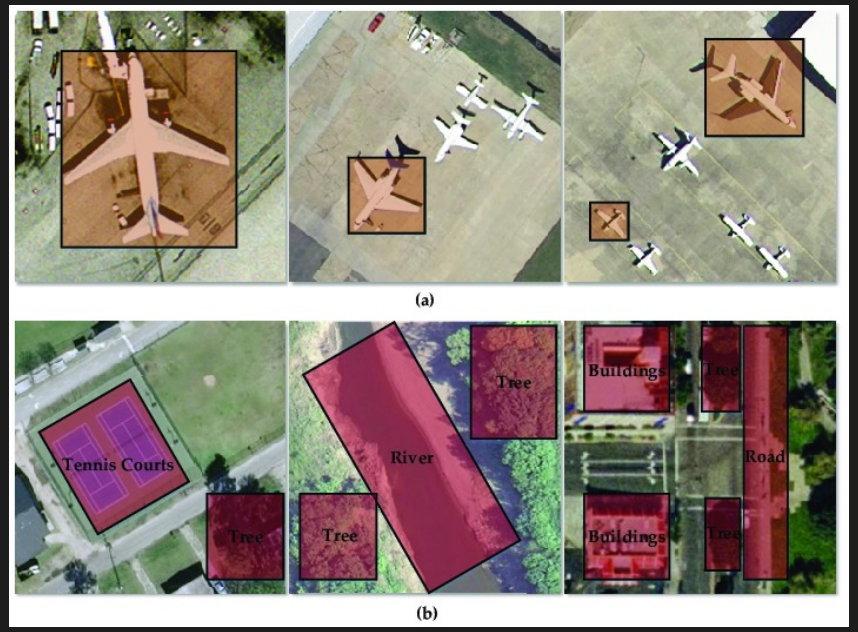
- Segmentation model from Google
- DeepLab, DeepLabv3, DeepLabv3+
- Rely on atrous convolution operations

DeepLab Family

Challenges in Semantic Segmentation

1. Enlarging receptive fields,
   - Deeper architecture, more abstract feature, small feature maps
2. Multi-scale objects

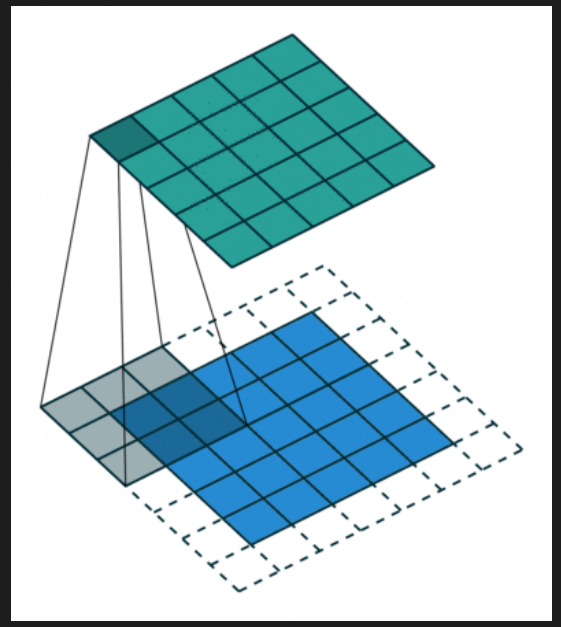


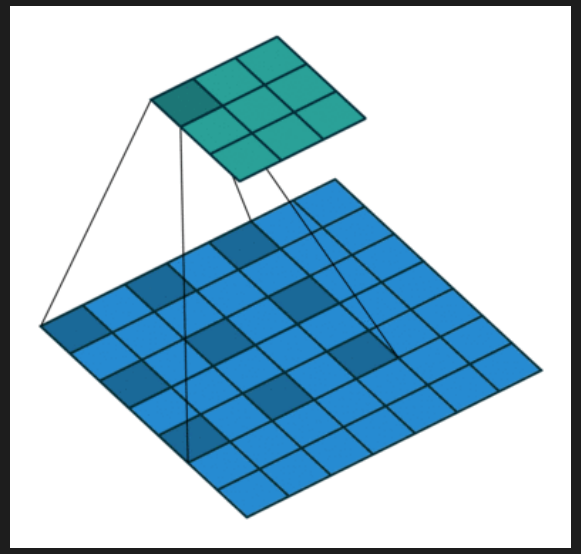


DeepLab Family

Atrous (Dilated) Convolution

- Dilation rate r
- Inject r-1 zeros inside a filter
- Larger receptive field with the same number of parameters and the same number of computations

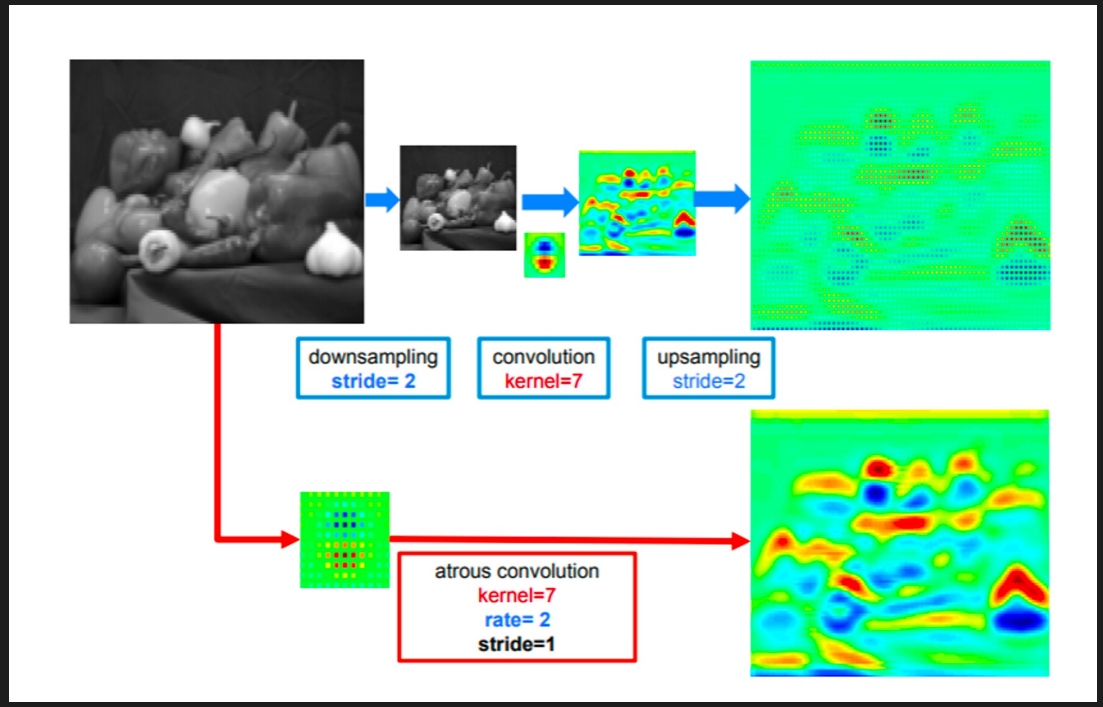




DeepLab Family

Atrous (Dilated) Convolution

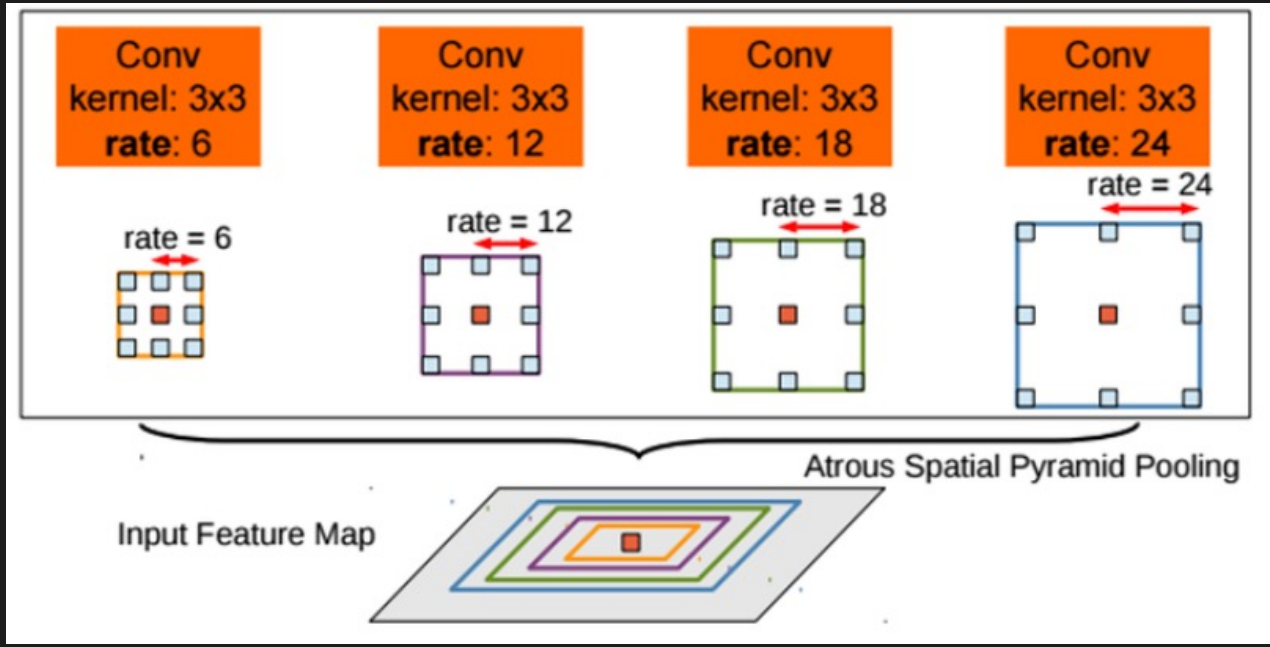
- No pooling; No lost spatial information
- Less information loss, better image resolution when upsampling



DeepLab Family

Atrous Spatial Pyramid Pooling (ASPP)

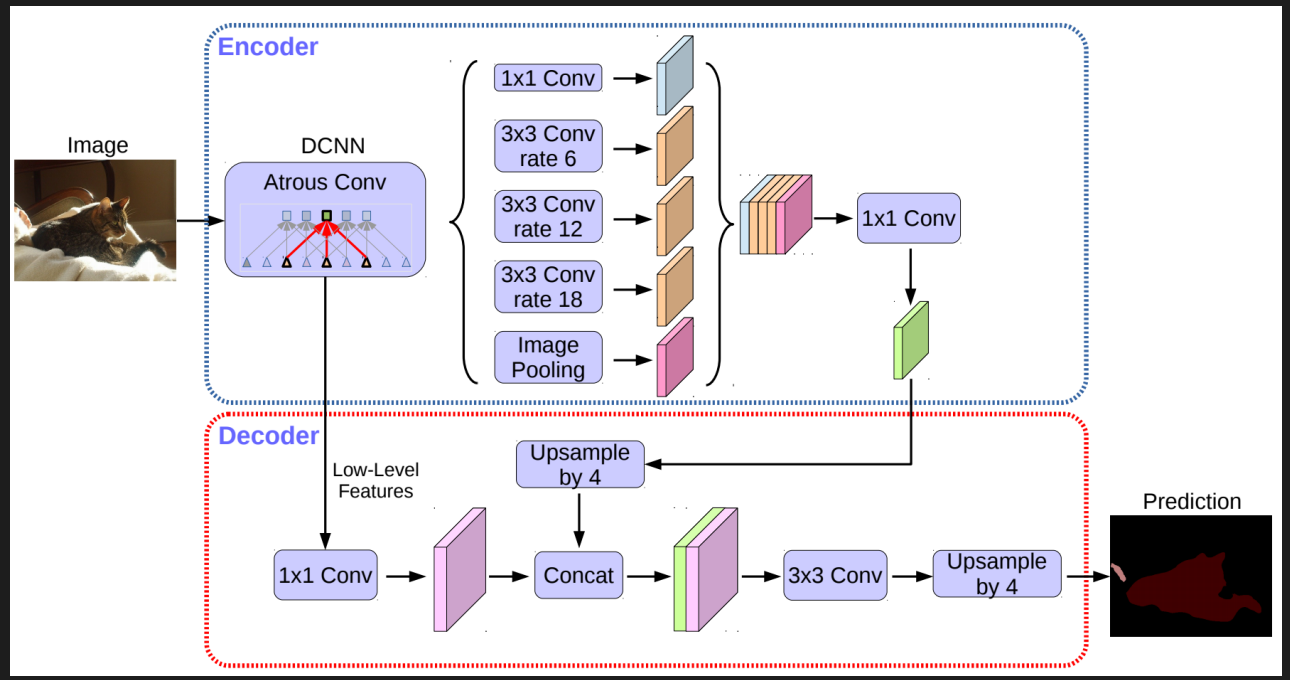
- Four parallel atrous convolutions with different dilation rates
- Each atrous convolution have different receptive field



DeepLab Family

Encoder-Decoder Architecture

- Encoder: atrous convolution, ASPP
- Decoder: skip architecture, upsampling



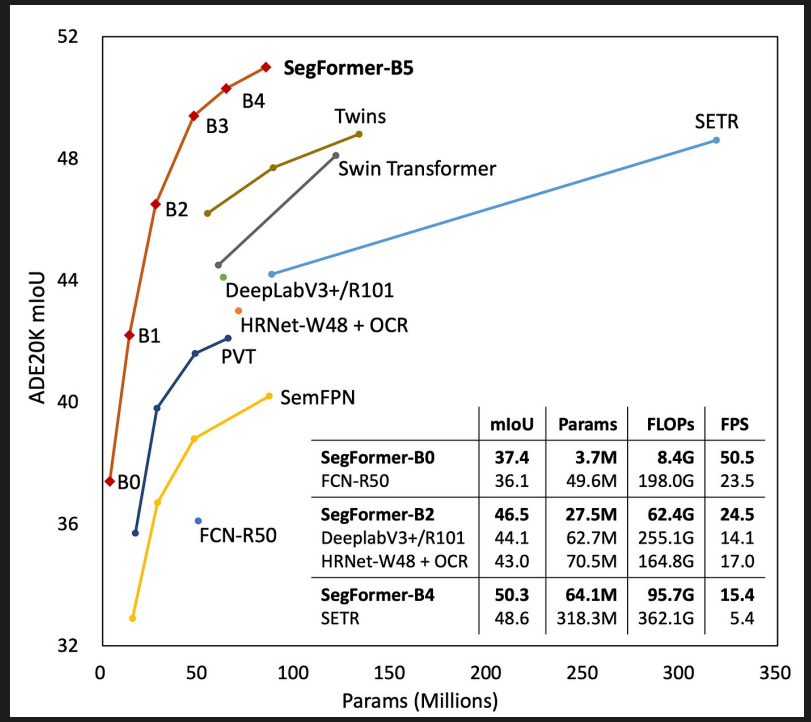
DeepLab Family: Summary

- Atrous convolution to efficiently enlarge a receptive field
- Catch multi-scale objects in a image through ASPP module
- Encoder-decoder architecture

SegFormer

SegFormer

- Transformer-based semantic segmentation model
- Simple and computationally efficient model
- Encoder-decoder architecture utilizing multi-level feature maps

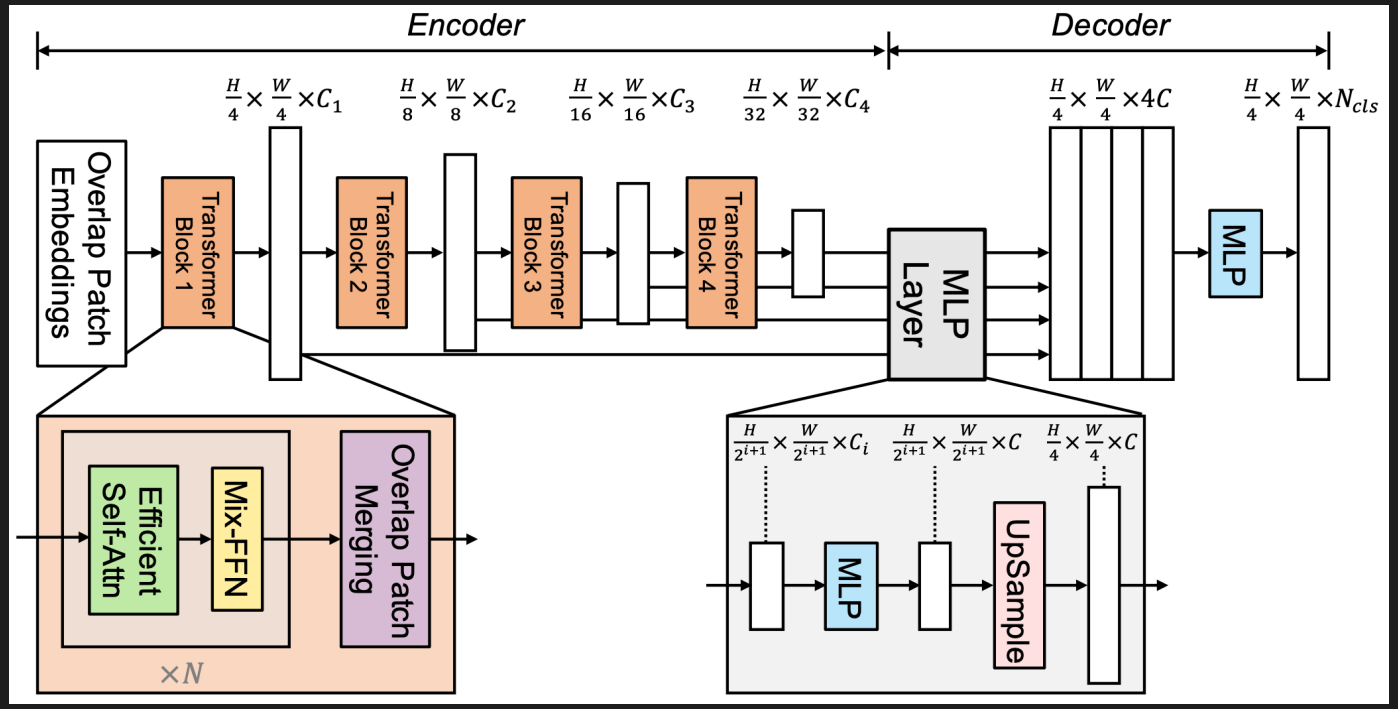


	mIoU	Params	FLOPs	FPS
SegFormer-B0	**37.4**	**3.7M**	**8.4G**	**50.5**
FCN-R50	36.1	49.6M	198.0G	23.5
SegFormer-B2	**46.5**	**27.5M**	**62.4G**	**24.5**
DeeplabV3+/R101	44.1	62.7M	255.1G	14.1
HRNet-W48 + OCR	43.0	70.5M	164.8G	17.0
SegFormer-B4	**50.3**	**64.1M**	**95.7G**	**15.4**
SETR	48.6	318.3M	362.1G	5.4

SegFormer

Hierarchical Transformer-based Encoder

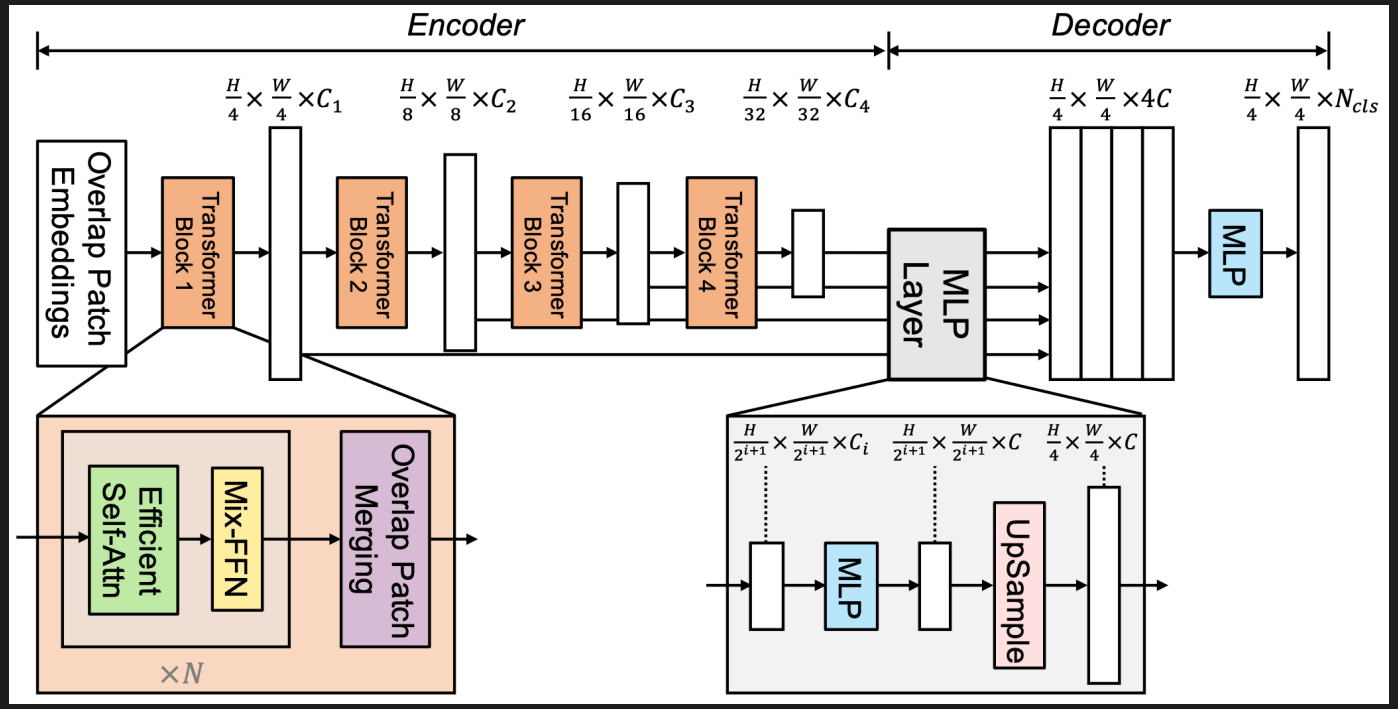
- Divide an image into 4X4 small patches for the dense prediction task
- Utilize feature maps from various layers



SegFormer

Lightweight All-MLP Decoder

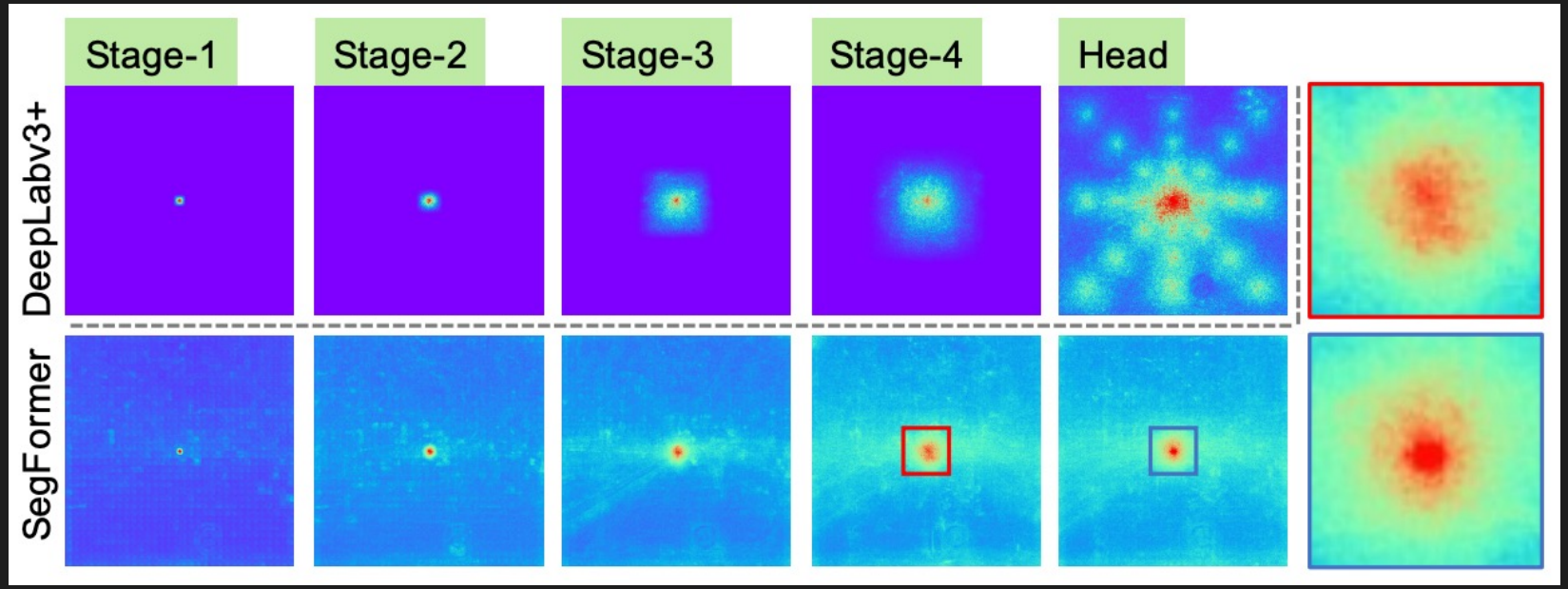
- Avoid hand-crafted and computationally demanding components in the others
- Since the encoder has a large receptive field, a simple decoder is enough



SegFormer

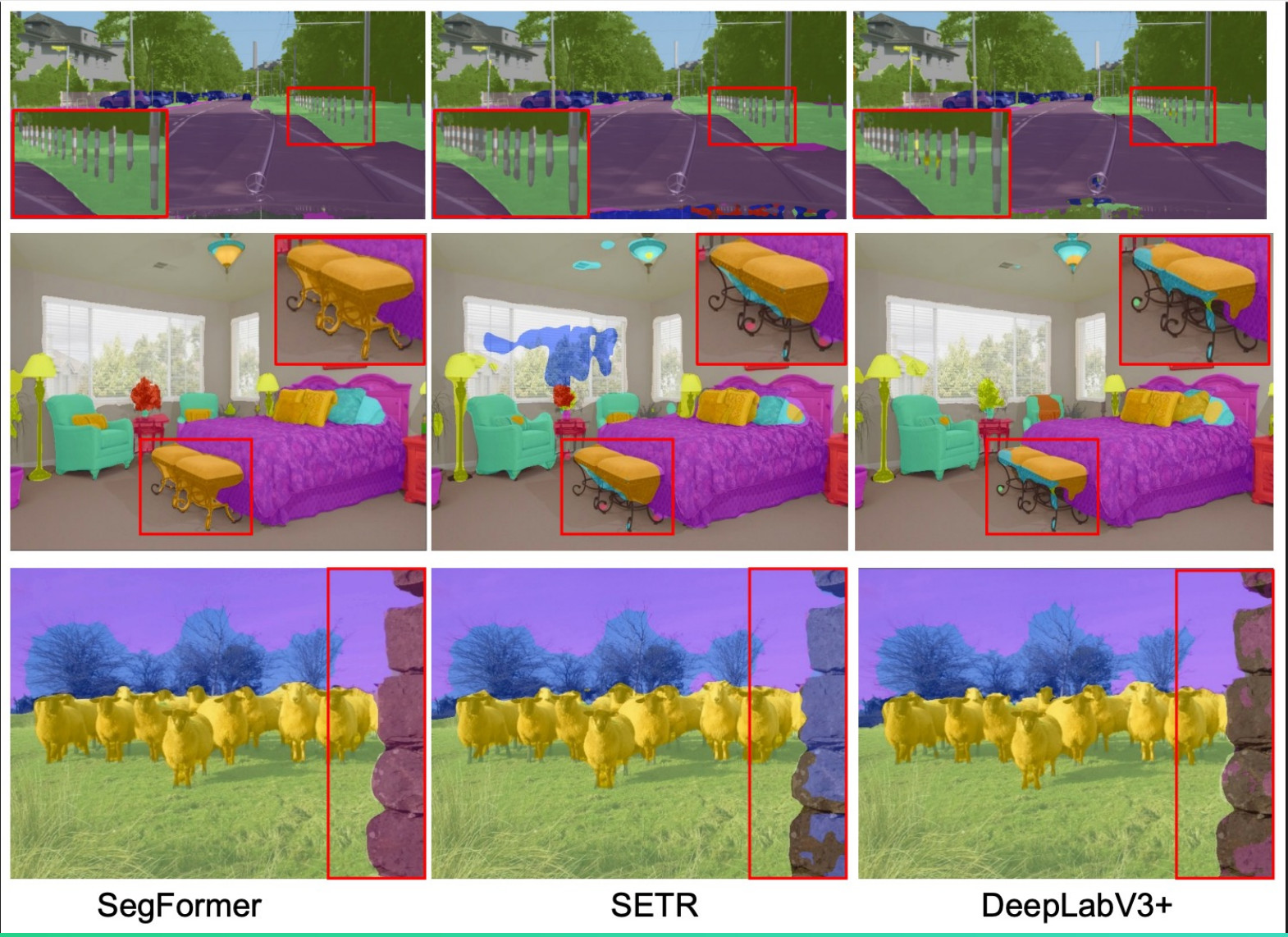
Effective Receptive Field Analysis

- The ERF of DeepLabv3+ is relatively small even at Stage-4, the deepest stage
- From local attentions to high-level context



SegFormer

Results



SegFormer

Results

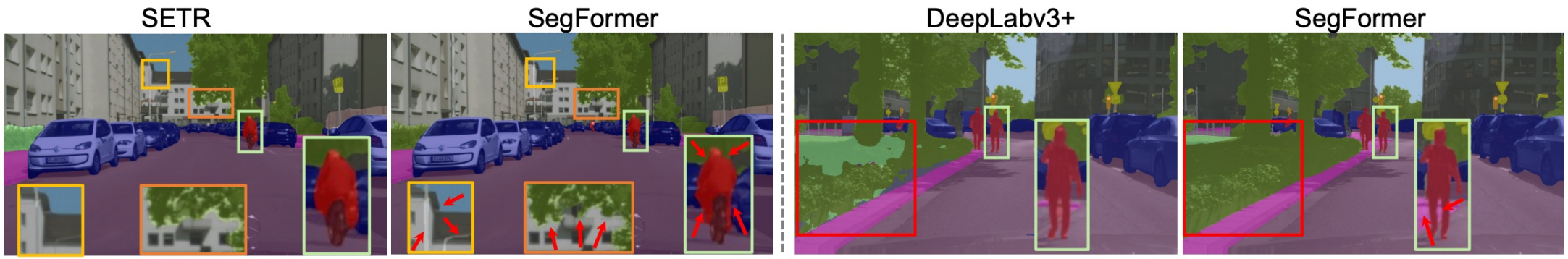


Figure 4: **Qualitative results on Cityscapes.** Compared to SETR, our SegFormer predicts masks with substantially finer details near object boundaries. Compared to DeeplabV3+, SegFormer reduces long-range errors as highlighted in red. Best viewed in screen.

SegFormer: Summary

- Hierarchical Transformer-based encoder
- Large effective receptive field
- Simple MLP Decoder

Others

Timeline of Segmentation Models

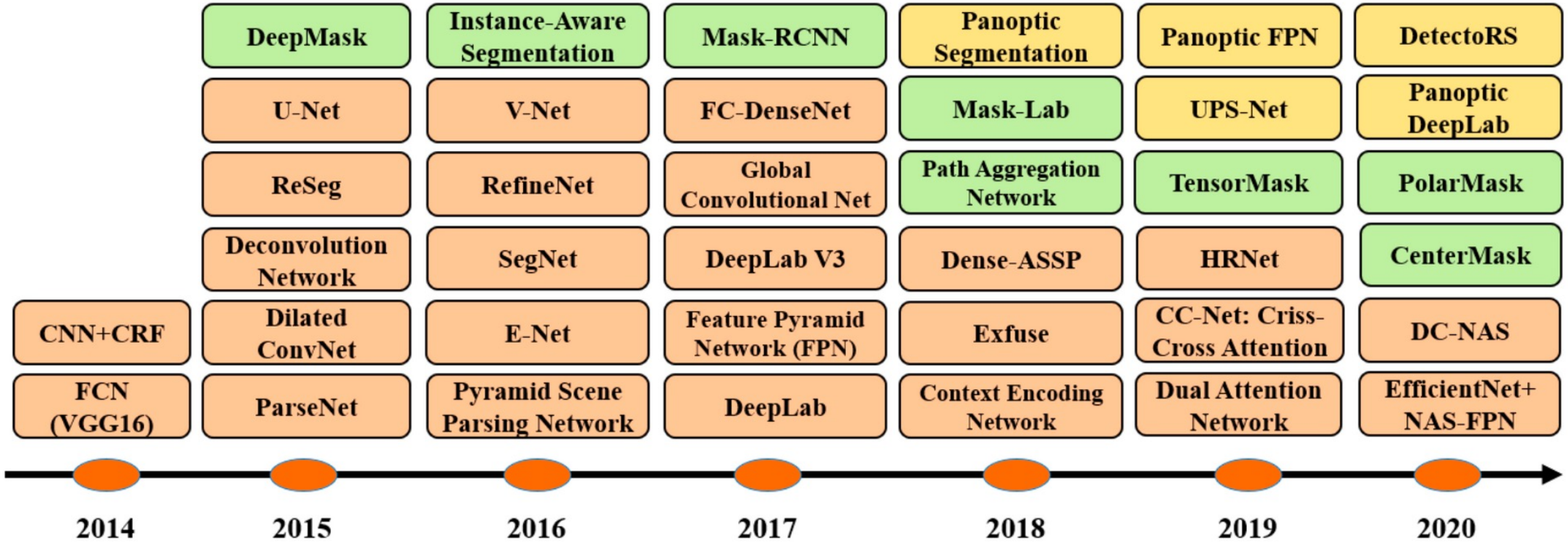


Fig. 32. The timeline of DL-based segmentation algorithms for 2D images, from 2014 to 2020. Orange, green, andn yellow blocks refer to semantic, instance, and panoptic segmentation algorithms respectively.

Segmentation Performance

TABLE 1
Accuracies of segmentation models on the PASCAL VOC test set. (* Refers to the model pre-trained on another dataset (such as MS-COCO, ImageNet, or JFT-300M).)

Method	Backbone	mIoU
FCN [31]	VGG-16	62.2
CRF-RNN [39]	-	72.0
CRF-RNN* [39]	-	74.7
BoxSup* [117]	-	75.1
Piecewise* [40]	-	78.0
DPN* [41]	-	77.5
DeepLab-CRF [78]	ResNet-101	79.7
GCN* [118]	ResNet-152	82.2
RefineNet [115]	ResNet-152	84.2
Wide ResNet [119]	WideResNet-38	84.9
PSPNet [56]	ResNet-101	85.4
DeeplabV3 [12]	ResNet-101	85.7
PSANet [98]	ResNet-101	85.7
EncNet [114]	ResNet-101	85.9
DFN* [99]	ResNet-101	86.2
Exfuse [120]	ResNet-101	86.2
SDN* [45]	DenseNet-161	86.6
DIS [123]	ResNet-101	86.8
DM-Net* [58]	ResNet-101	87.06
APC-Net* [60]	ResNet-101	87.1
EMANet [95]	ResNet-101	87.7
DeeplabV3+ [83]	Xception-71	87.8
Exfuse [120]	ResNeXt-131	87.9
MSCI [61]	ResNet-152	88.0
EMANet [95]	ResNet-152	88.2
DeeplabV3+* [83]	Xception-71	89.0
EfficientNet+NAS-FPN [135]	-	90.5

TABLE 2
Accuracies of segmentation models on the Cityescapes dataset.

Method	Backbone	mIoU
FCN-8s [31]	-	65.3
DPN [41]	-	66.8
Dilation10 [79]	-	67.1
DeeplabV2 [78]	ResNet-101	70.4
RefineNet [115]	ResNet-101	73.6
FoveaNet [124]	ResNet-101	74.1
Ladder DenseNet [125]	Ladder DenseNet-169	73.7
GCN [118]	ResNet-101	76.9
DUC-HDC [80]	ResNet-101	77.6
Wide ResNet [119]	WideResNet-38	78.4
PSPNet [56]	ResNet-101	85.4
BiSeNet [126]	ResNet-101	78.9
DFN [99]	ResNet-101	79.3
PSANet [98]	ResNet-101	80.1
DenseASPP [81]	DenseNet-161	80.6
SPGNet [127]	2xResNet-50	81.1
DANet [93]	ResNet-101	81.5
CCNet [96]	ResNet-101	81.4
DeeplabV3 [12]	ResNet-101	81.3
AC-Net [129]	ResNet-101	82.3
OCR [44]	ResNet-101	82.4
GS-CNN [128]	WideResNet	82.8
HRNetV2+OCR (w/ASPP) [44]	HRNetV2-W48	83.7
Hierarchical MSA [137]	HRNet-OCR	85.1

Other Models

- SegNet, HRNet, Mask R-CNN
- (Transformer-based) MaskFormer, Segmenter
- (Trends) SAM and its variants

Further Challenges

Further Challenges

Challenging Datasets

- A large number of objects and overlapping objects

Interpretability

Weakly-Supervised and Unsupervised Learning

Memory Efficient Models

- Available in edge devices

3D Point-Cloud Segmentation

- Unordered and unstructured data

Discussion Assignment

8월 8일 오후 4시까지 노션에 제출하세요 ☺

Discussion Assignment 1

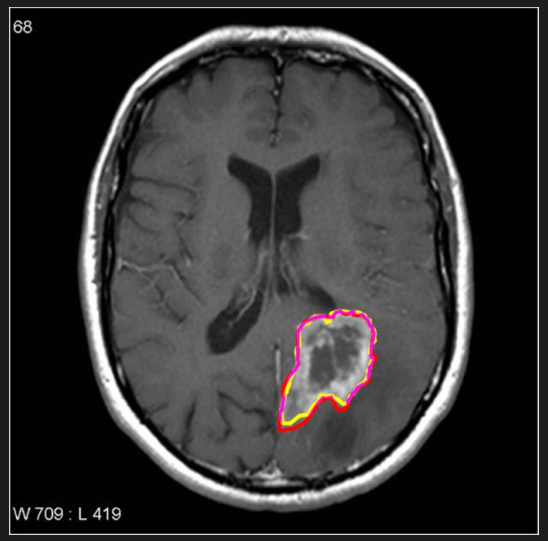
Try to “Segment Anything!”

- https://segment-anything.com/demo
- 데모 사이트에서 최신 세그멘테이션 모델 SAM을 이용해보기
- 최소 하나 이상의 이미지에 세그멘테이션을 시도한 후 스크린샷과 소감 제출하기

Discussion Assignment 2

How to Deal with Unbalanced Classes

- 실제 이미지에서 물체가 차지하는 영역이 작을 경우, 배경만 많이 맞춰도 성능이 잘 나온다
- 의료 이미지의 경우 물체를 정확하게 찾는 것이 중요한데, 이러한 문제를 어떻게 해결할 수 있을까?



Programming Assignment 3

8월 15일 자정까지 노션에 제출하세요 ☺

Programming Assignment 3

- Due Date: 2023.08.15. 23:59 Github Classroom Link
- Colab 활용
- 주제 : Implementing Object Detector based on YOLO and Faster R-CNN
- ‘TODO’ 부분 코드 추가하기 (epoch 수 제외 코드 수정 금지)
- TODO 부분만 읽고 구현하지 말고, 처음부터 꼼꼼히 파일 전체를 읽으면서 흐름을 이해하며 과제를 진행하는 것을 적극 추천
- Discussion 질문에 대한 답변도 필수!
- 학습에 꽤 오랜 시간을 투자해야 함 -> 반드시 미리미리 하는 것을 적극 권장!!
- 질문이 있다면 언제든지 자유롭게 “배민성”에게 연락주시면 됩니다!

참고자료

자료 출처

- Image Segmentation Using Deep Learning: A Survey
- U-Net: Convolutional Networks for Biomedical Image Segmentation
- DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs
- Rethinking Atrous Convolution for Semantic Image Segmentation
- Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation
- SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers
- FCN 논문 리뷰 — Fully Convolutional Networks for Semantic Segmentation

감사합니다.