Image Segmentation

Naeemullah Khan
naeemullah.khan@kaust.edu.sa

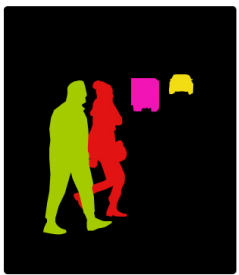
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

LMH
Lady Margaret Hall

July 25, 2025

Types of Image Segmentation

SEMANTIC IMAGE SEGMENTATION



INSTANCE SEGMENTATION

PANOPTIC SEGMENTATION

Table of Contents

Table of Contents (cont.)

Image Segmentation: **Motivation**

Why Segment Images?

- **Classification**: Tells *what* is in the image.
- **Detection**: Tells *where* objects are.
- **Segmentation**: Tells *exactly which pixels* belong to each object.

Crucial for:

- Medical imaging
- Autonomous driving
- Robotics
- AR/VR
- Satellite imagery

Real-World Impact

- Tumor localization in MRIs
- Road boundary detection for self-driving cars
- Crop field segmentation from drone images

Image Segmentation: **Learning Outcomes**

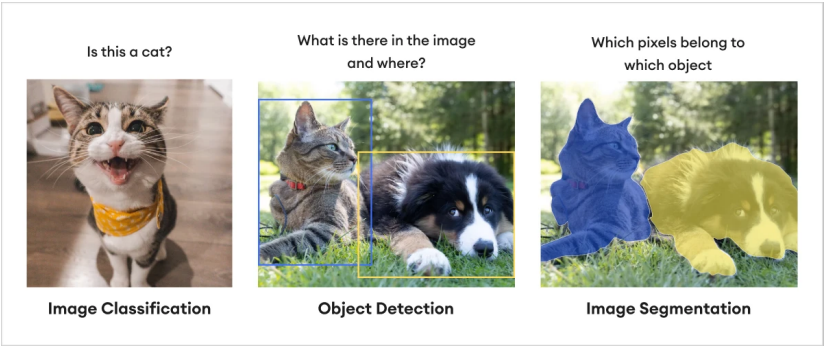
Learning Outcomes

- Understand the need and core idea behind image segmentation
- Explain transposed convolution, Fully Convolutional Networks (FCNs), and U-Net
- Differentiate between semantic, instance, and panoptic segmentation
- Identify key challenges and limitations in image segmentation
- Explore future directions and advanced research topics

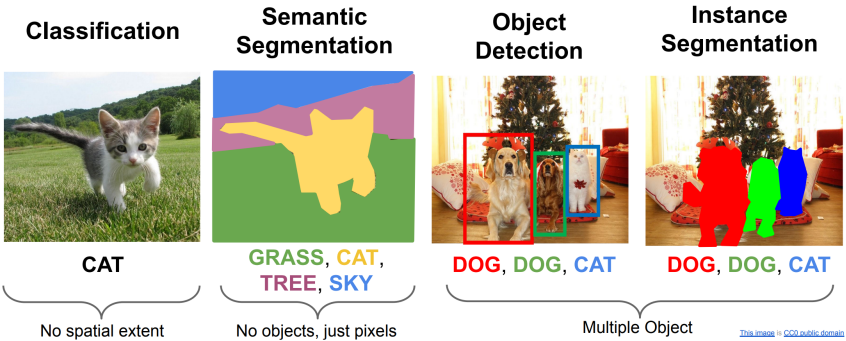
Image Segmentation: **Introduction**

What is Image Segmentation?

- **Definition:** Partitioning an image into multiple meaningful segments.
- **Goal:** Assign a label to every pixel such that pixels with the same label share visual characteristics.



Computer Vision Tasks



- **Classification:** Identify the main object in an image.
- **Detection:** Locate objects with bounding boxes.
- **Segmentation:** Classify each pixel to delineate object boundaries.

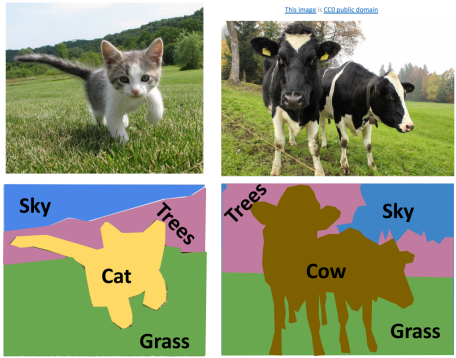
Challenges in Segmentation

- **Pixel-level labeling is labor-intensive**
- Varying object size and shape
- Occlusion and overlapping objects
- Class imbalance (background vs. object)

Image Segmentation: **Things and Stuff**

Things and Stuff

- **Things:** Object categories that can be separated into object instances (e.g. cats, cars, person)
- **Stuff:** Object categories that cannot be separated into instances (e.g. sky, grass, water, trees)



Computer Vision Tasks

- **Object Detection:** Detects individual object instances, but only gives box(Only things!)

- **Semantic Segmentation:** Gives per-pixel labels, but merges instances (Both things and stuff)

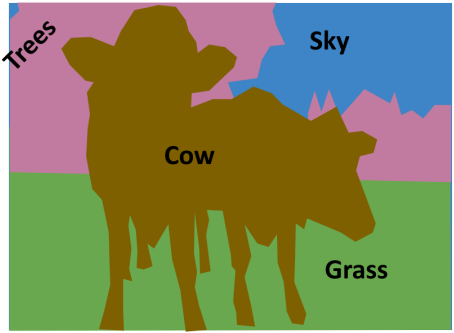
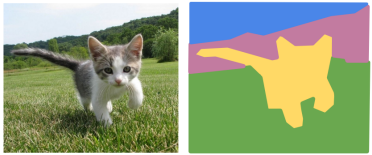


Image Segmentation: **Semantic Segmentation**

Semantic Segmentation



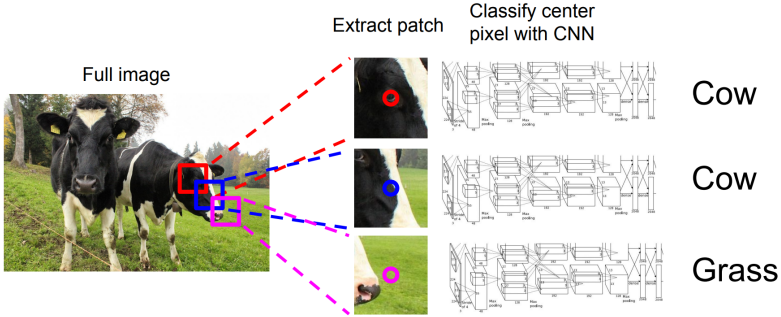
GRASS, CAT,
TREE, SKY, ...

Paired training data: for each training image,
each pixel is labeled with a semantic category.

At test time, classify each pixel of a new image.

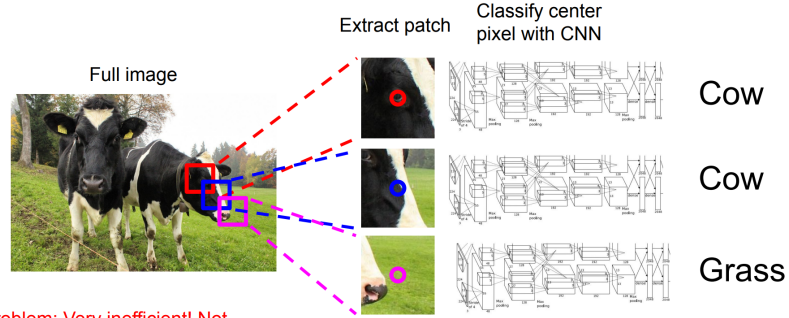
Semantic Segmentation (cont.)

Semantic Segmentation Idea: Sliding Window



Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation Idea: Sliding Window (cont.)

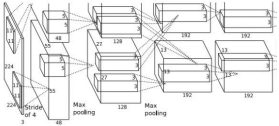


Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

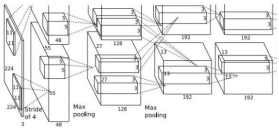
Semantic Segmentation Idea: Convolution

Full image



An intuitive idea: encode the entire image with conv net, and do semantic segmentation on top.

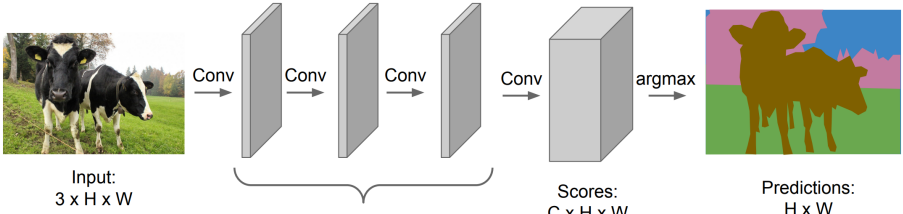
Full image



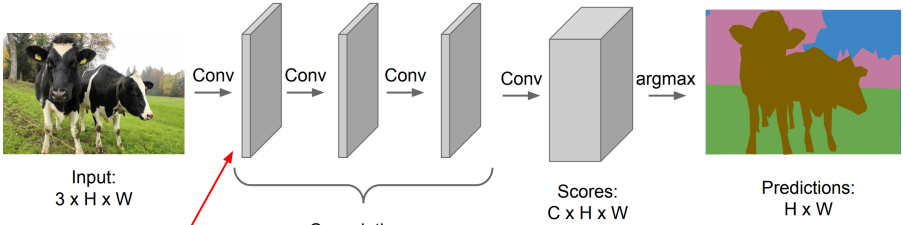
An intuitive idea: encode the entire image with conv net, and do semantic segmentation on top.

Problem: classification architectures often reduce feature spatial sizes to go deeper, but semantic segmentation requires the output size to be the same as input size.

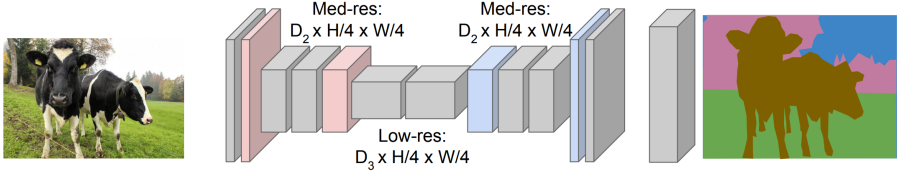
Design a network with only convolutional layers without downsampling operators to make predictions for pixels all at once!



Design a network with only convolutional layers without downsampling operators to make predictions for pixels all at once!



Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!



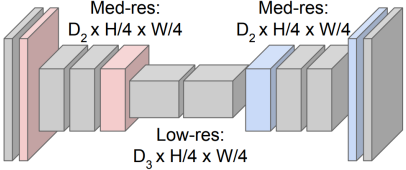
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: ???

Input: 3 x H x W



High-res: D_1 x H/2 x W/2

High-res: D_1 x H/2 x W/2

C x H x W

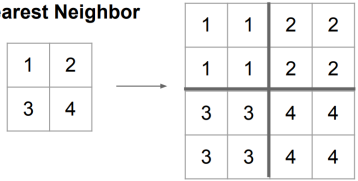
Predictions: H x W

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Image Segmentation: **In-Network Upsampling**

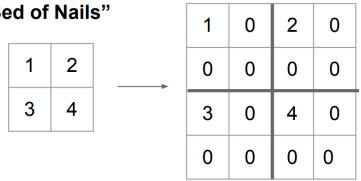
In-Network Upsampling: Unpooling

Nearest Neighbor



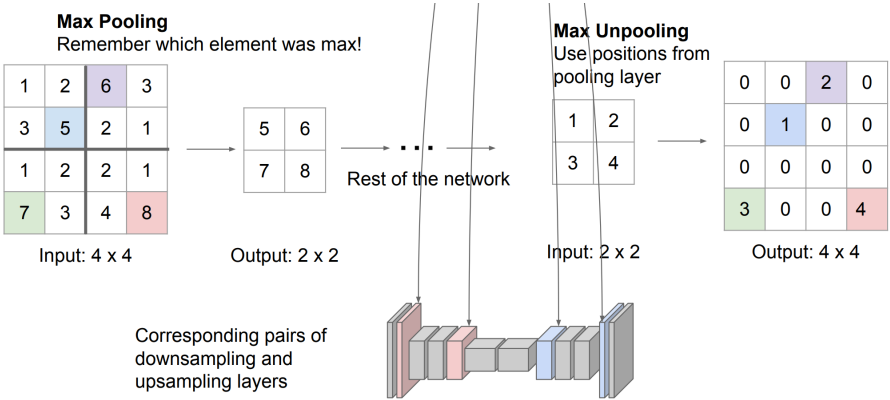
Input: 2 x 2 Output: 4 x 4

“Bed of Nails”

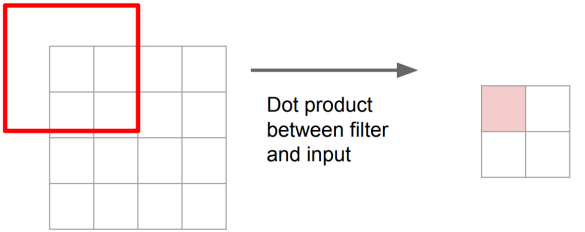


Input: 2 x 2 Output: 4 x 4

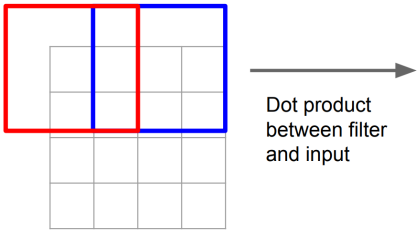
In-Network Upsampling: Max Unpooling



Recall: Normal 3 x 3 convolution, stride 2 pad 1



Recall: Normal 3 x 3 convolution, stride 2 pad 1

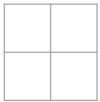


Filter moves 2 pixels in the input for every one pixel in the output

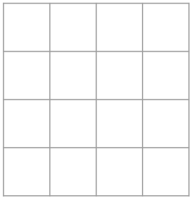
Stride gives ratio between movement in input and output

We can interpret strided convolution as “learnable downsampling”.

3 x 3 **transposed** convolution, stride 2 pad 1

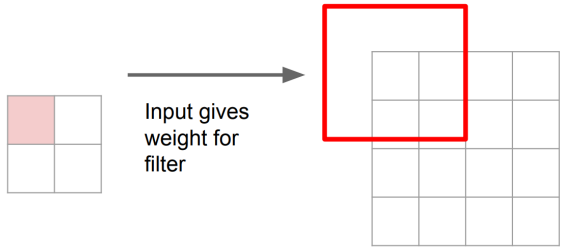


Input: 2 x 2

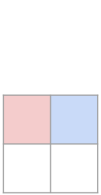


Output: 4 x 4

3 x 3 **transposed** convolution, stride 2 pad 1

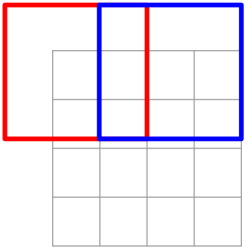


3 x 3 **transposed** convolution, stride 2 pad 1



Input: 2 x 2

Input gives weight for filter



Output: 4 x 4

Filter moves 2 pixels in the output for every one pixel in the input

Stride gives ratio between movement in output and input

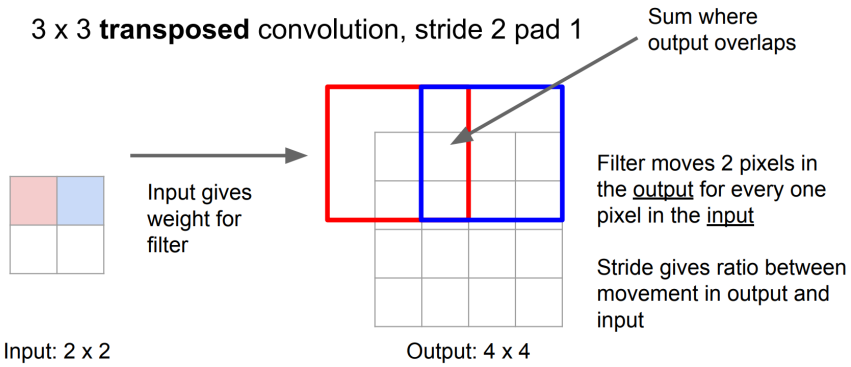
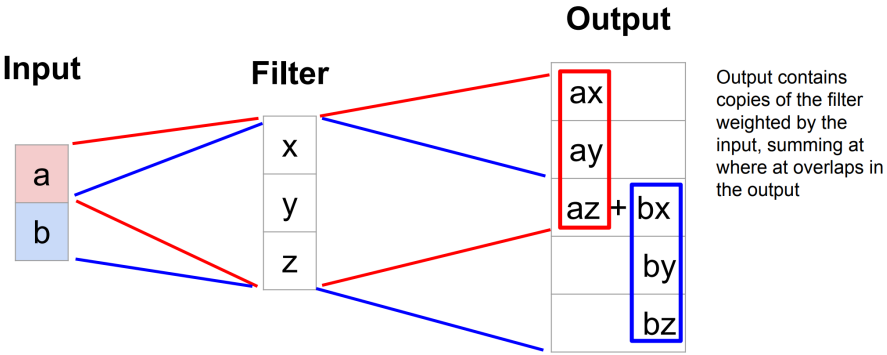

3 x 3 transposed convolution, stride 2 pad 1
Sum where output overlaps
Input gives weight for filter
Filter moves 2 pixels in the output for every one pixel in the input
Stride gives ratio between movement in output and input
Input: 2 x 2
Output: 4 x 4

Image Segmentation: **Transposed Convolution**

Transposed Convolution: 1D Example



Output contains copies of the filter weighted by the input, summing at where at overlaps in the output

Image Segmentation: **Fully Convolutional Networks (FCNs)**

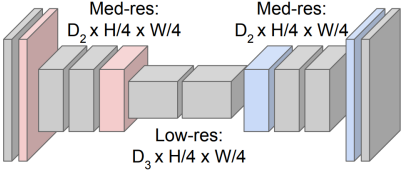
Semantic Segmentation Idea: Fully Convolutional

Downsampling:
Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling:
Unpooling or strided transposed convolution

Input:
3 x H x W



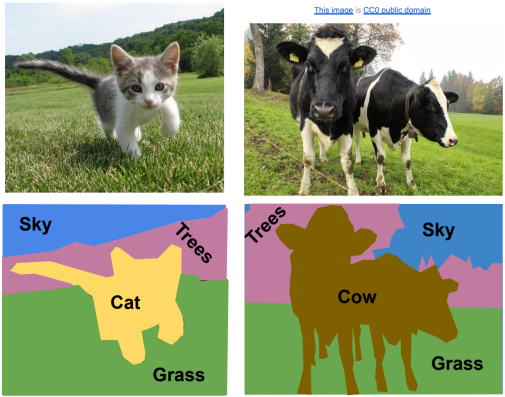
High-res:
D_1 x H/2 x W/2

High-res:
D_1 x H/2 x W/2

Predictions:
H x W

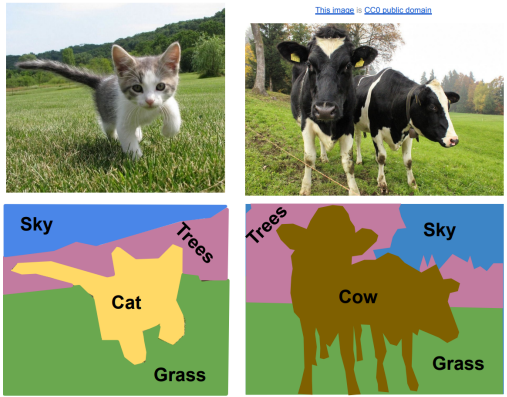
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

- Label each pixel in the image with a category label



Semantic Segmentation

- Label each pixel in the image with a category label

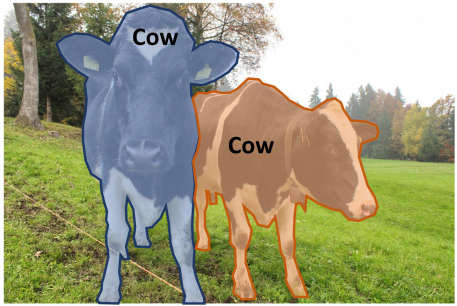


- **Does not differentiate instances, only care about pixels**

Image Segmentation: **Instance Segmentation**

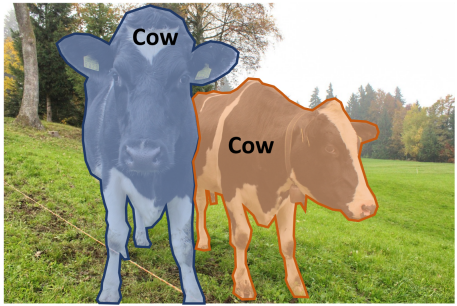
Instance Segmentation

- Detect all objects in the image, and identify the pixels that belong to each object (Only things!)



Instance Segmentation

- Detect all objects in the image, and identify the pixels that belong to each object (Only things!)
- **Approach:** Perform object detection, then predict a segmentation mask for each object!

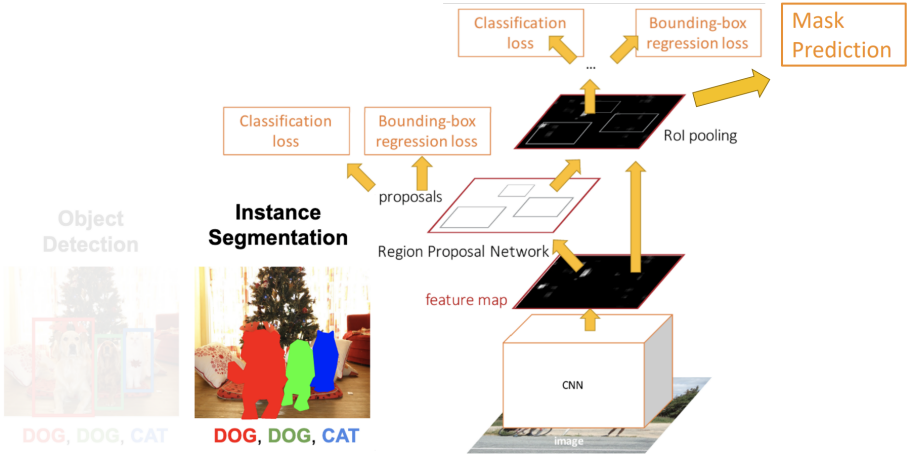


Instance Segmentation: **Mask R-CNN**

Mask R-CNN Overview

- **Developed on top of Faster R-CNN**
- Faster R-CNN outputs:
  - Class label
  - Bounding-box offset
- Mask R-CNN adds a third branch:
  - Predicts object mask for each candidate
- Performs both **Semantic** and **Instance Segmentation**

Mask R-CNN (cont.)



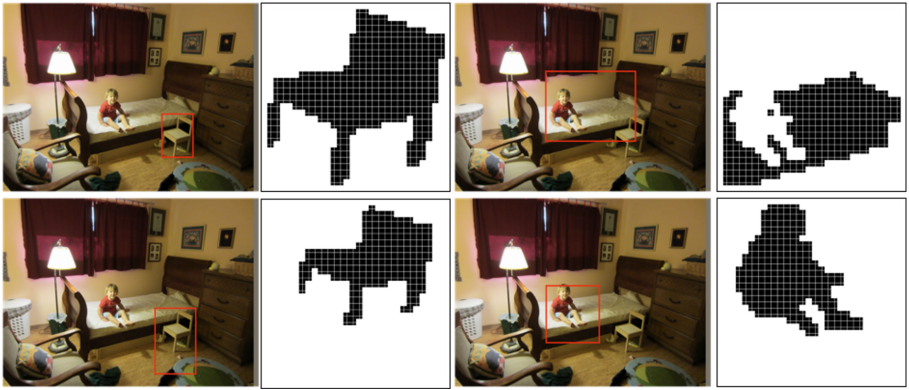
Advantages of Mask R-CNN

- **Simplicity**: Simple to train
- **Performance**: Outperforms previous methods, works almost in real-time
- **Efficiency**: Very efficient, small overhead compared to Faster R-CNN
- **Flexibility**: Can perform detection and estimation tasks simultaneously

[0]He et al, "Mask R-CNN", ICCV 2017

Mask R-CNN: Example Training Targets

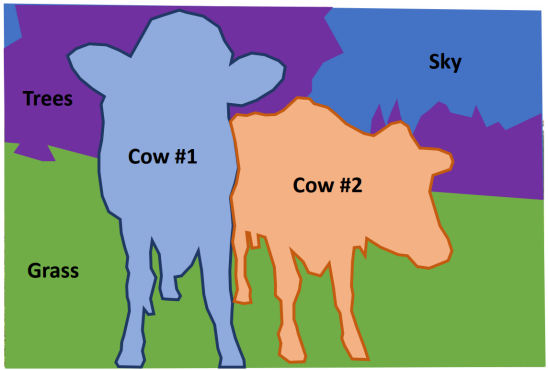
Mask R-CNN: Example Training Targets



Mask R-CNN: Very Good Results!

Image Segmentation: **Panoptic Segmentation**

- Label all pixels in the image (both things and stuff)
- For "thing" categories also separate into instances

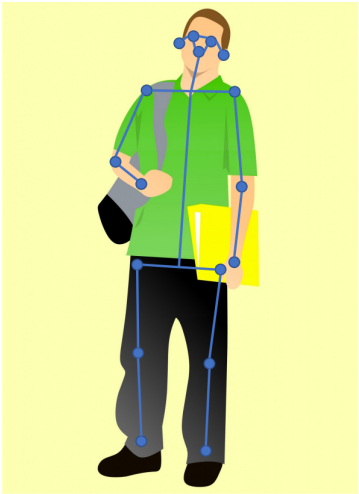


Beyond Instance Segmentation: Panoptic Segmentation

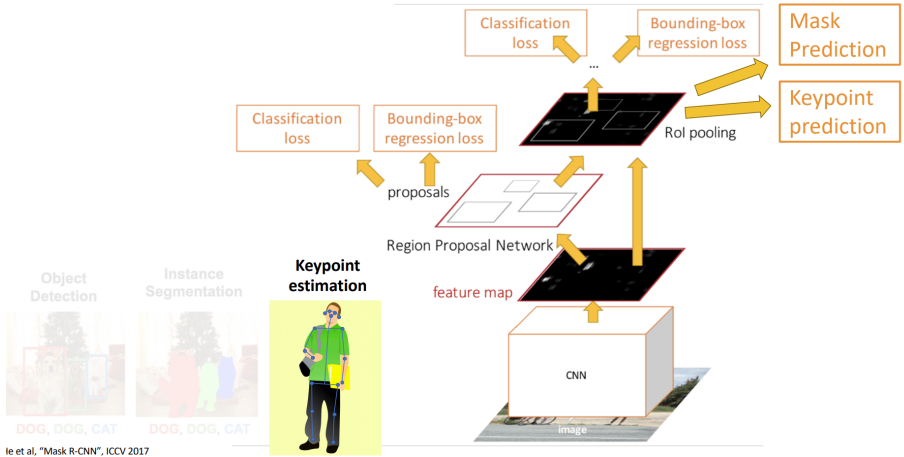
[0]Kirillov et al, "Panoptic Feature Pyramid Networks", CVPR 2019

Image Segmentation: **Human Keypoints Estimation**

- Represent the pose of a human by locating a set of keypoint e.g. 17 keypoints:
- Nose
- Left / Right eye
- Left / Right earLeft / Right shoulder
- Left / Right elbow
- Left / Right wrist



Mask R-CNN: Keypoint Estimation



Ie et al, "Mask R-CNN", ICCV 2017

Mask R-CNN: Keypoint Estimation

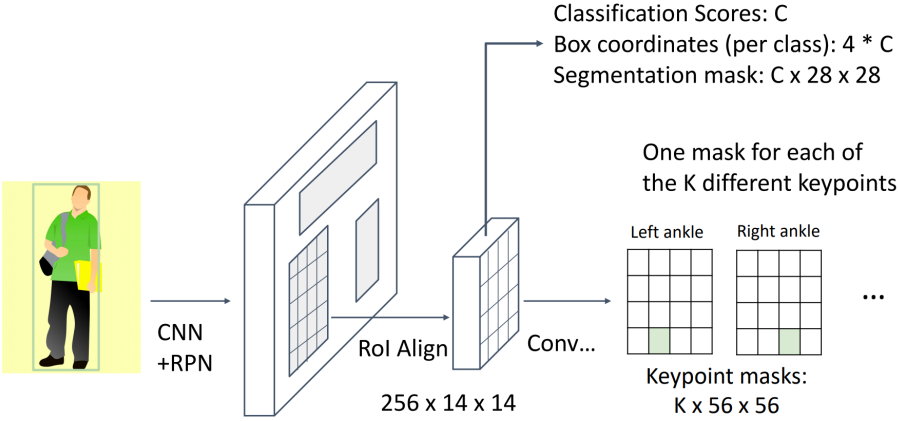
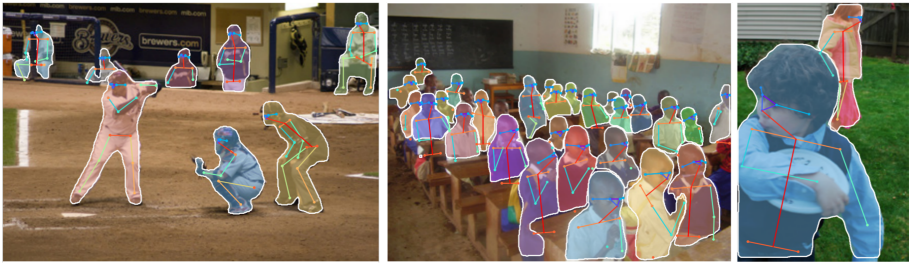


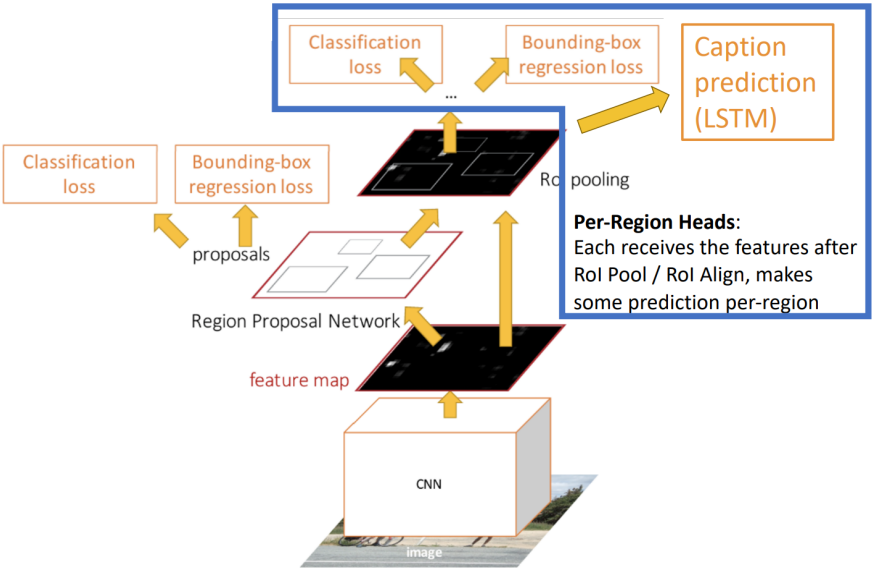
Image Segmentation: **Pose Estimation**



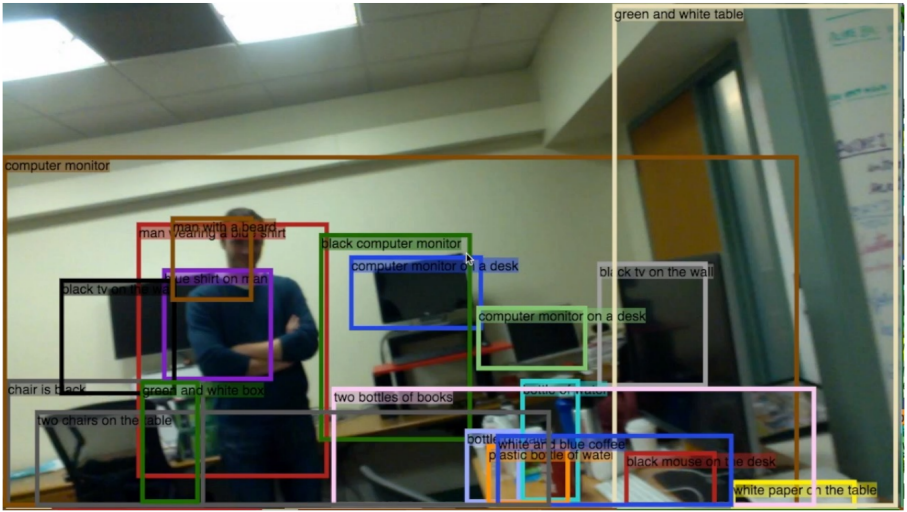
[0]He et al, "Mask R-CNN", ICCV 2017

Image Segmentation: **Captioning**

Captioning: Predict a caption per region!



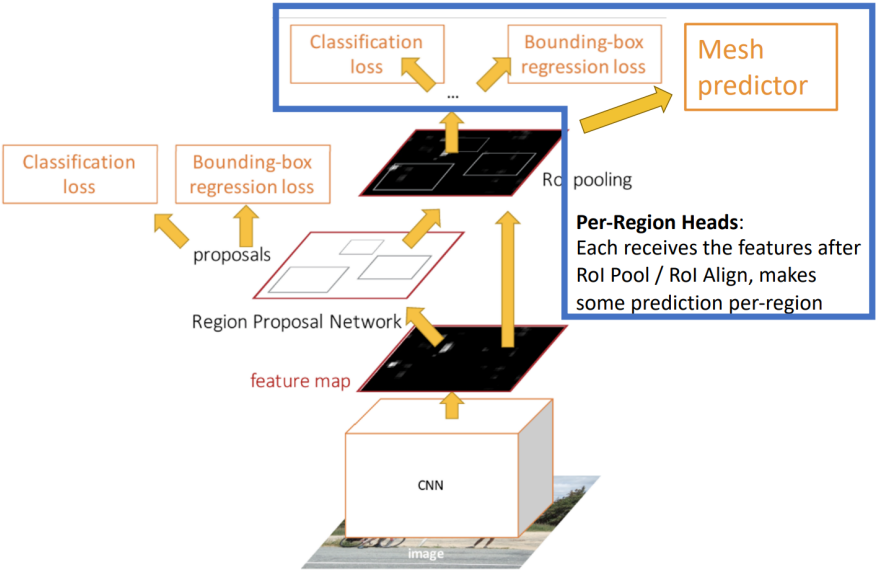
Captioning: Predict a caption per region!



Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization

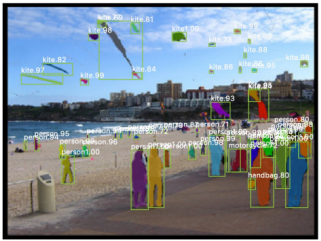
Image Segmentation: **3D Shape Prediction**

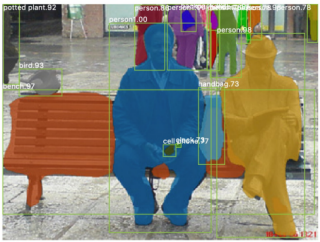
3D Shape Prediction



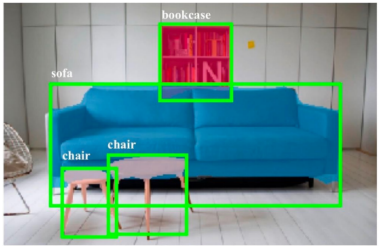
3D Shape Prediction

Mask R-CNN:
2D Image -> 2D shapes





Mesh R-CNN:
2D Image -> **3D** shapes



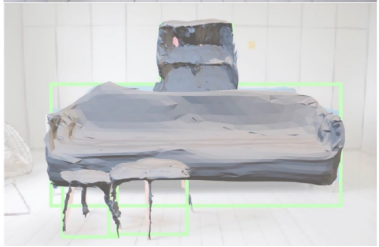


Image Segmentation: **Object Tracking**

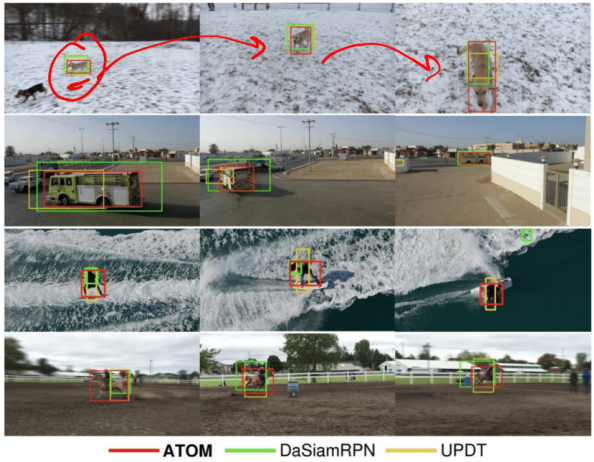
Goal of Object Tracking

- Track objects over a sequence of photos or a video
- Exceedingly challenging in multi-object tracking scenarios
- Need to ensure objects are not mixed up or lost midway

A Common Solution

- Perform object detection in each frame
- Assign unique IDs to each detected object
- Store feature vectors for each object (e.g., appearance, position)
- Track objects across frames based on their IDs and feature vectors

A Common Solution (cont.)



Comparison of 3 approaches for Object Tracking

Image Segmentation: **Limitations and Future Directions**

Common Limitations

- Requires a large amount of labeled data (pixel-wise annotations are expensive)
- Computationally intensive
- Struggles with:
  - Occlusion
  - Small objects
  - Rare classes
- Poor generalization across domains (e.g., synthetic $\rightarrow$ real)

Where is the Field Going?

- **Self-supervised learning:** Reduce need for labeled data
- **Few-shot segmentation:** Learn from few examples
- **Transformer-based segmentation:** (e.g., Segmenter, Mask2Former)
- **Real-time segmentation:** For AR/VR, autonomous systems
- **Multimodal segmentation:** Combine RGB + depth + text

Image Segmentation: **References**

References & Further Reading

[1] Long, J., Shelhamer, E., & Darrell, T. (2015).

Fully Convolutional Networks for Semantic Segmentation.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

arxiv.org/abs/1411.4038

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015).

U-Net: Convolutional Networks for Biomedical Image Segmentation.

In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

arxiv.org/abs/1505.04597

References & Further Reading (cont.)

[3] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017).

Mask R-CNN.

In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

arxiv.org/abs/1703.06870

[4] Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019).

Panoptic Segmentation.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

arxiv.org/abs/1801.00868

References & Further Reading (cont.)

[5] Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021).

Segmenter: Transformer for Semantic Segmentation.

In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

arxiv.org/abs/2105.05633

[6] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2022).

SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers.

In *Advances in Neural Information Processing Systems (NeurIPS)*.

arxiv.org/abs/2105.15203

Additional Resources:

- PapersWithCode: paperswithcode.com/task/image-segmentation
- Stanford CS231n Slides (Segmentation Modules): cs231n.stanford.edu/slides/2022/
- DeepMind Perceiver IO: arxiv.org/abs/2107.14795
- Google Mask2Former: arxiv.org/abs/2112.01527

Credits

Dr. Prashant Aparajeya

Computer Vision Scientist — Director(AISimply Ltd)

p.aparajeya@aisimply.uk

This project benefited from external collaboration, and we acknowledge their contribution with gratitude.