Computer Vision

Naeemullah Khan

naeemullah.khan@kaust.edu.sa

KAUST Academy
King Abdullah University of Science and Technology

January 07, 2024

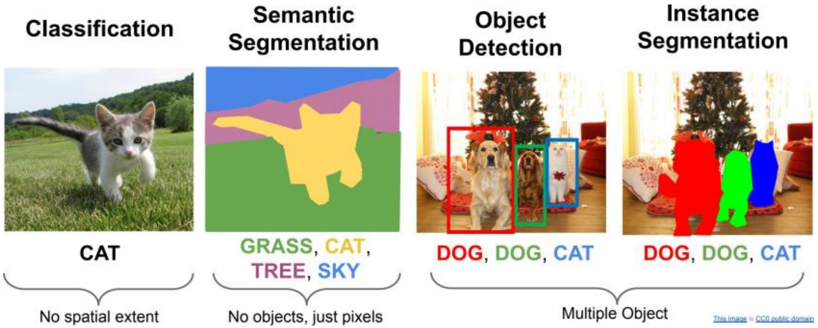
Image Classification

- Previously, we discussed Image Classification
- A core task in Computer Vision

This image by Nikita is licensed under CC-BY 2.0

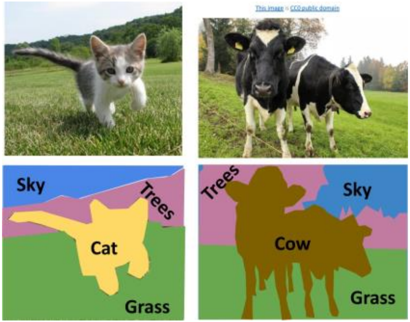
(assume given a set of possible labels)
{dog, cat, truck, plane, ...}

→ cat


Classification
Semantic Segmentation
Object Detection
Instance Segmentation
CAT
GRASS, CAT, TREE, SKY
DOG, DOG, CAT
DOG, DOG, CAT
No spatial extent
No objects, just pixels
Multiple Object
This image is CC0 public domain

Things and Stuff

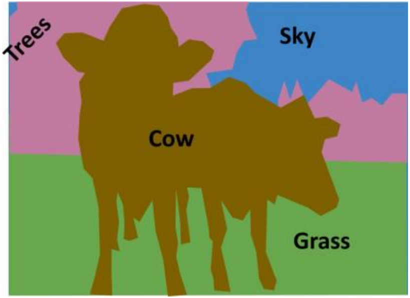
- **Things:** Object categories that can be separated into object instances (e.g. cats, cars, person)
- **Stuff:** Object categories that cannot be separated into instances (e.g. sky, grass, water, trees)



Computer Vision Tasks

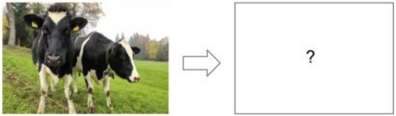
- **Object Detection:** Detects individual object instances, but only gives box(Only things!)

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

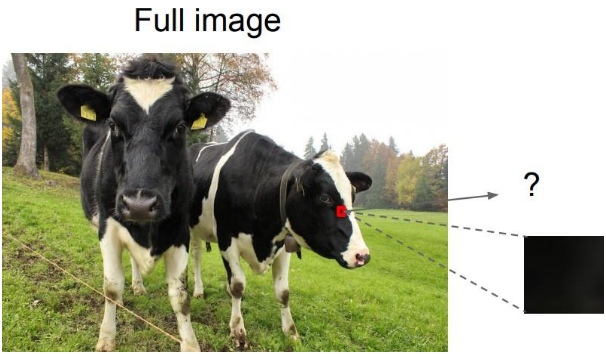


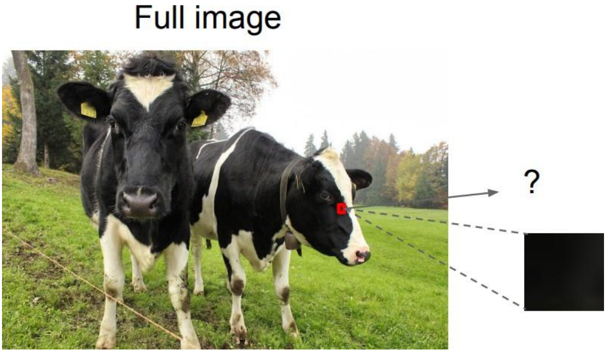
Semantic Segmentation

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



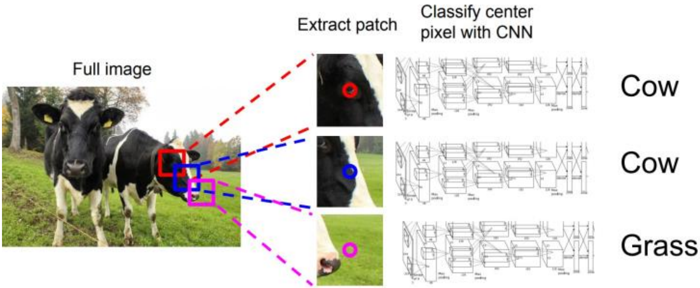
At test time, classify each pixel of a new image.


Full image
?



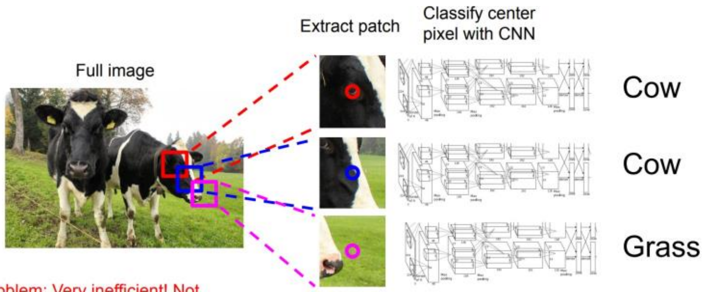
- Impossible to classify without context
- How do we include context?

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.)

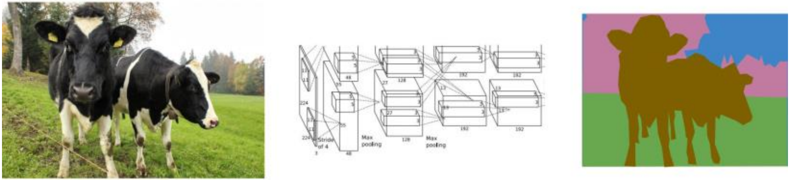


Problem: Very inefficient! Not reusing shared features between overlapping patches

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

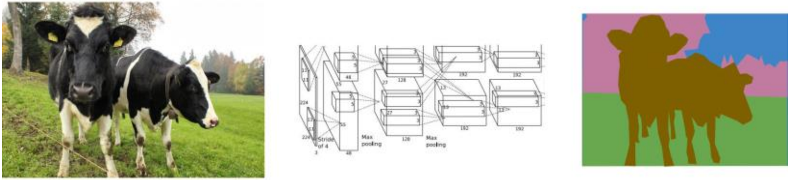
Full image



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

Semantic Segmentation Idea: Convolution (cont.)

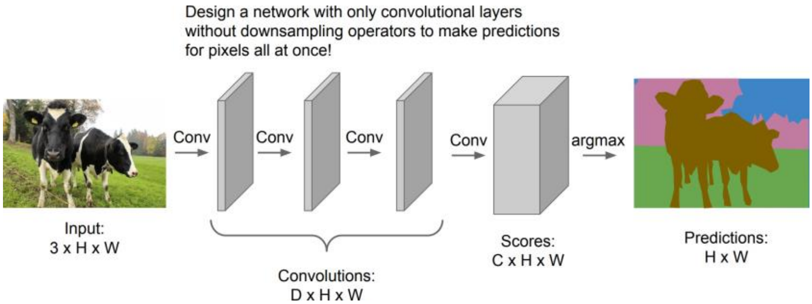
Full image



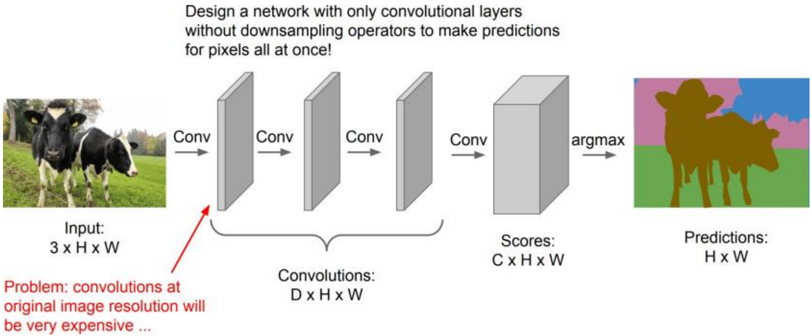
An intuitive idea: encode the entire image with conv net, and do semantic segmentation on top.

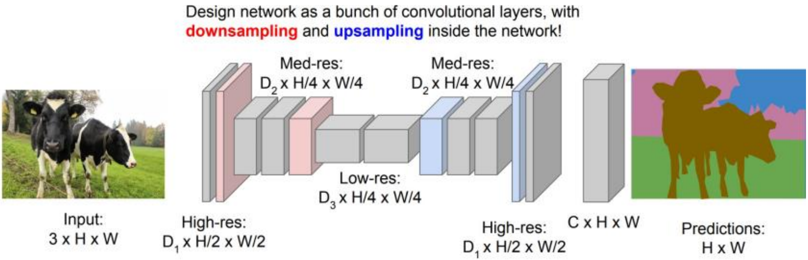
Problem: classification architectures often reduce feature spatial sizes to go deeper, but semantic segmentation requires the output size to be the same as input size.

Semantic Segmentation Idea: Fully Convolutional

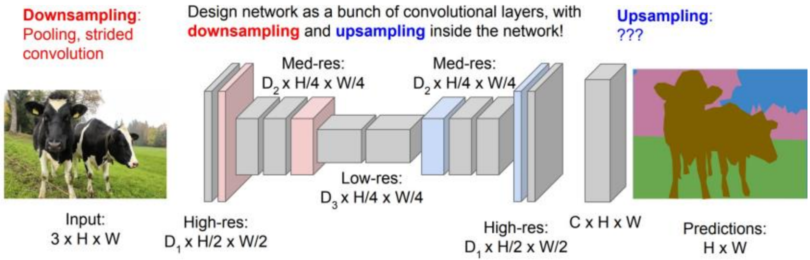


Semantic Segmentation Idea: Fully Convolutional (cont.)



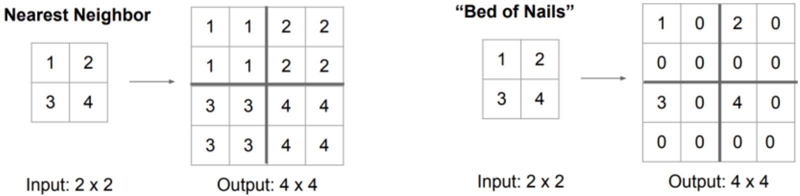


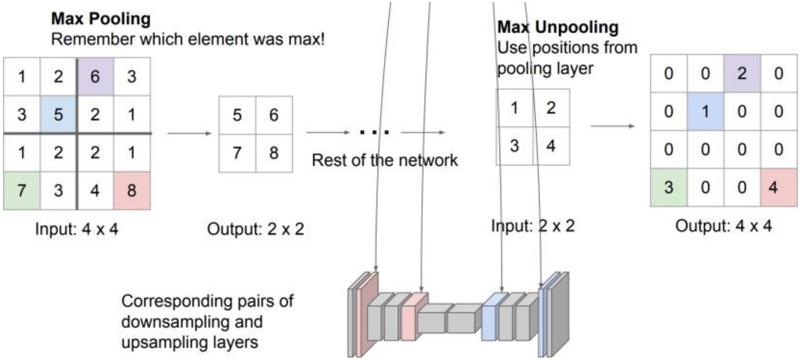
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015



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

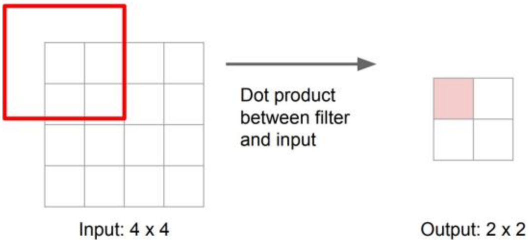
In-Network Upsampling: Unpooling



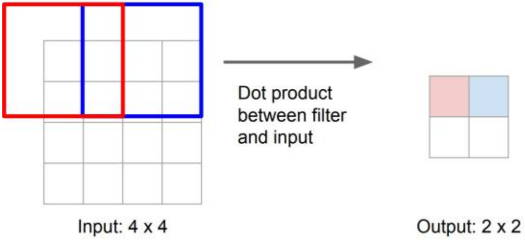

Max Pooling
Remember which element was max!
1 2 6 3
3 5 2 1
1 2 2 1
7 3 4 8
Input: 4 x 4
5 6
7 8
Output: 2 x 2
. . .
Rest of the network
Max Unpooling
Use positions from pooling layer
1 2
3 4
Input: 2 x 2
0 0 2 0
0 1 0 0
0 0 0 0
3 0 0 4
Output: 4 x 4
Corresponding pairs of downsampling and upsampling layers

Learnable Upsampling: Transposed Convolution

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



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



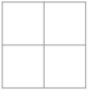
Filter moves 2 pixels in the input for every one pixel in the output

Stride gives ratio between movement in input and output

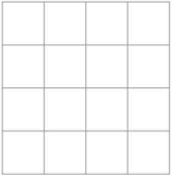
We can interpret strided convolution as "learnable downsampling".

Learnable Upsampling: Transposed Convolution (cont.)

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

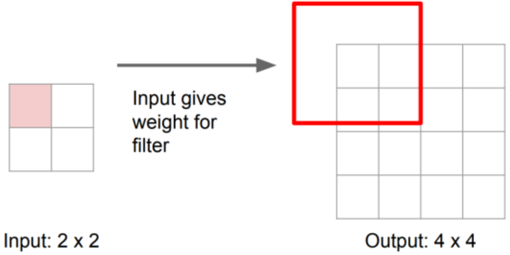


Input: 2 x 2

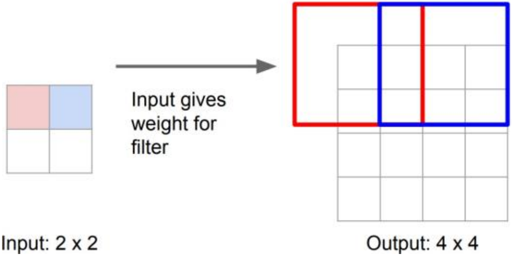


Output: 4 x 4

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



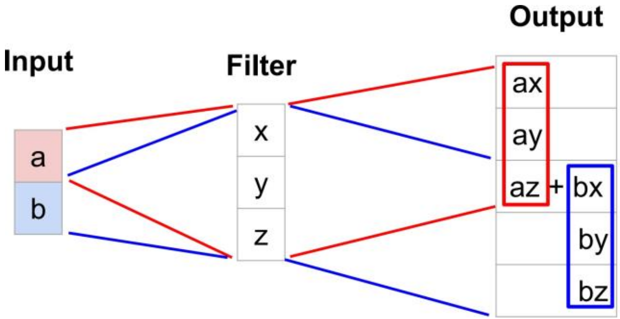
3 x 3 **transposed** convolution, stride 2 pad 1



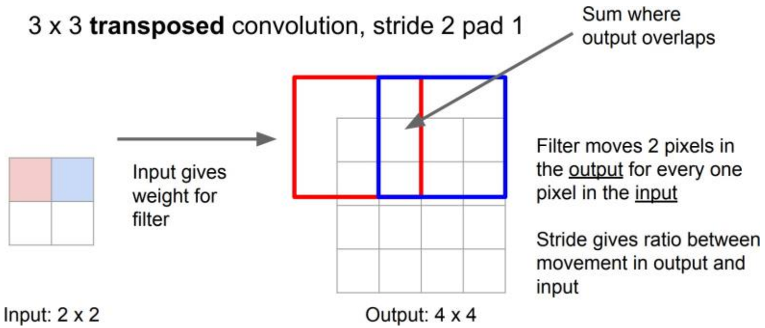
Filter moves 2 pixels in the output for every one pixel in the input

Stride gives ratio between movement in output and input

Transposed Convolution: 1D Example

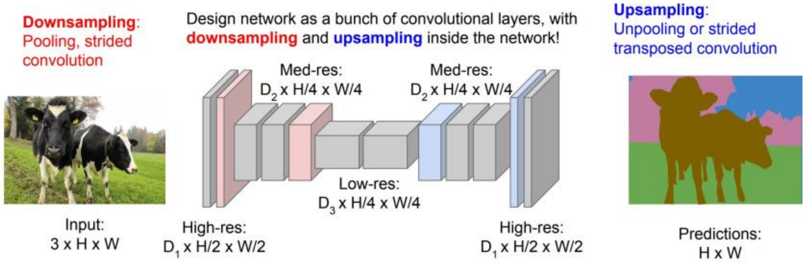


Output contains copies of the filter weighted by the input, summing at where at overlaps in the output



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

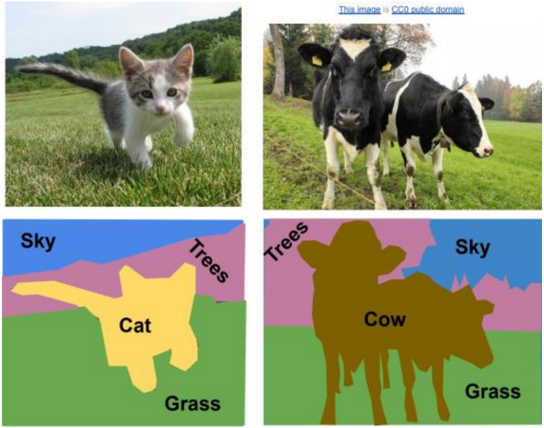
Semantic Segmentation Idea: Fully Convolutional



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

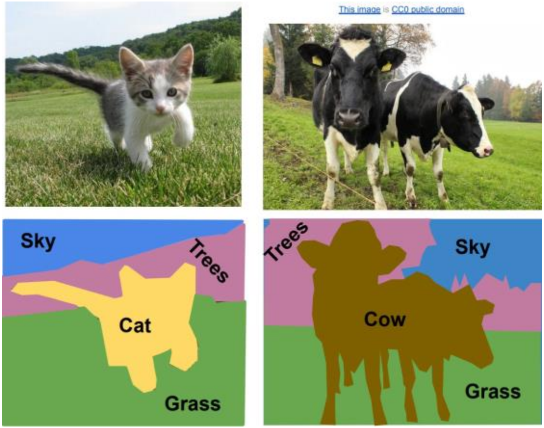
Semantic Segmentation

- Label each pixel in the image with a category label



Semantic Segmentation

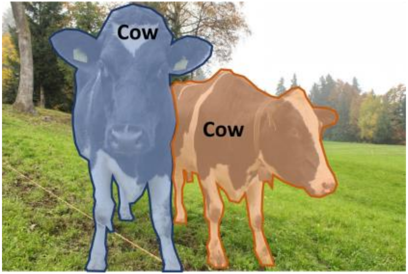
- Label each pixel in the image with a category label



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

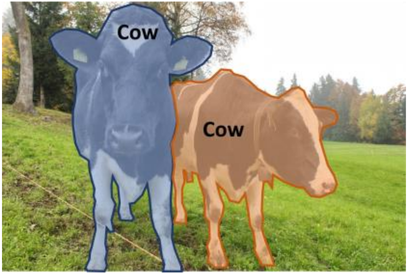
Instance Segmentation

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

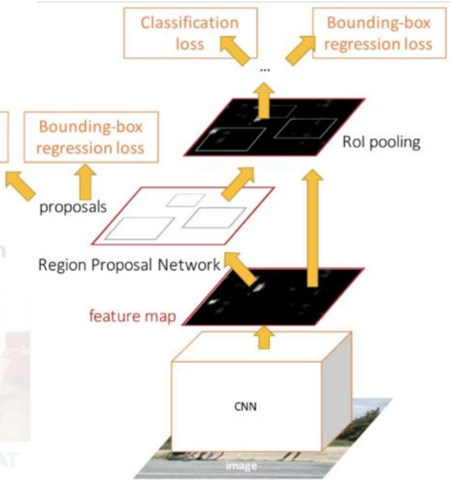


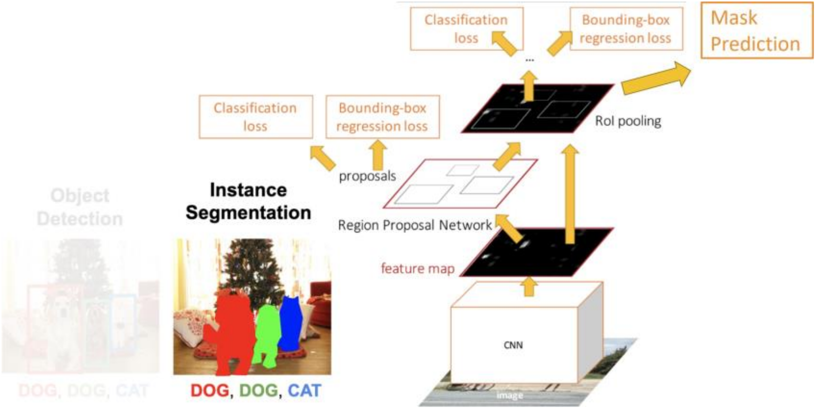
Instance Segmentation

- 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!



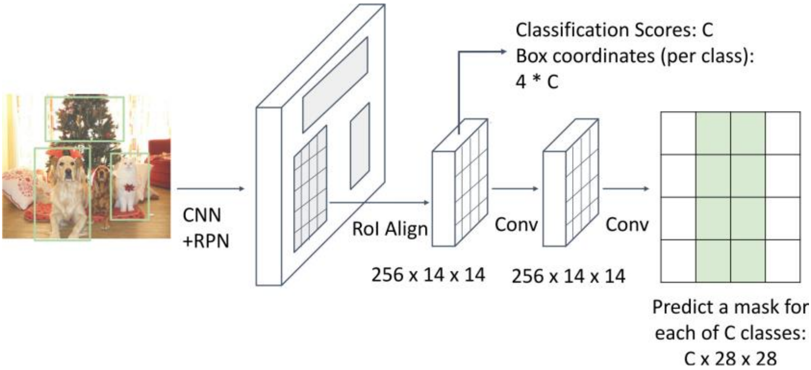
Object Detection: Faster R-CNN



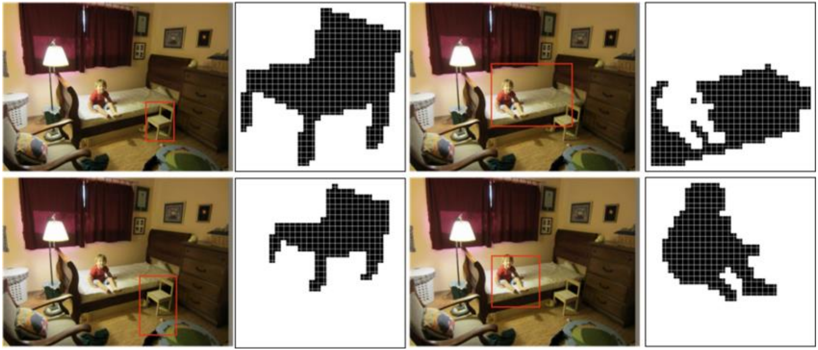


[0] He et al, "Mask R-CNN", ICCV 2017

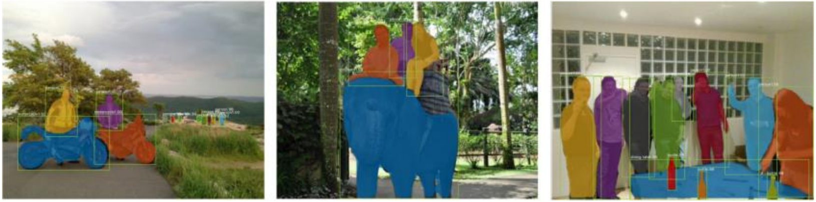
Instance Segmentation: Mask R-CNN



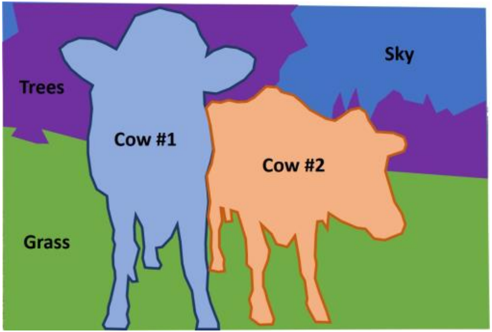
Mask R-CNN: Example Training Targets



Mask R-CNN: Very Good Results!



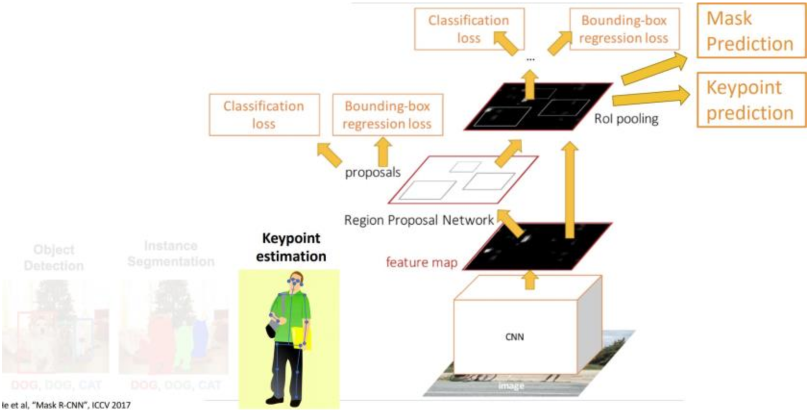
- Label all pixels in the image (both things and stuff)
- For "thing" categories also separate into instances



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

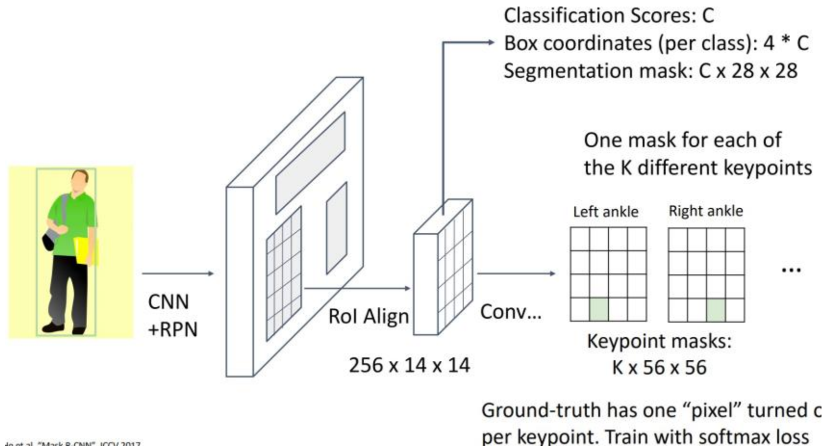
- Represent the pose of a human by locating a set of keypoint se.g. 17 keypoints:
- Nose
- Left / Right eye
- Left / Right earLeft / Right shoulder
- Left / Right elbow
- Left / Right wrist

Mask R-CNN: Keypoint Estimation



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

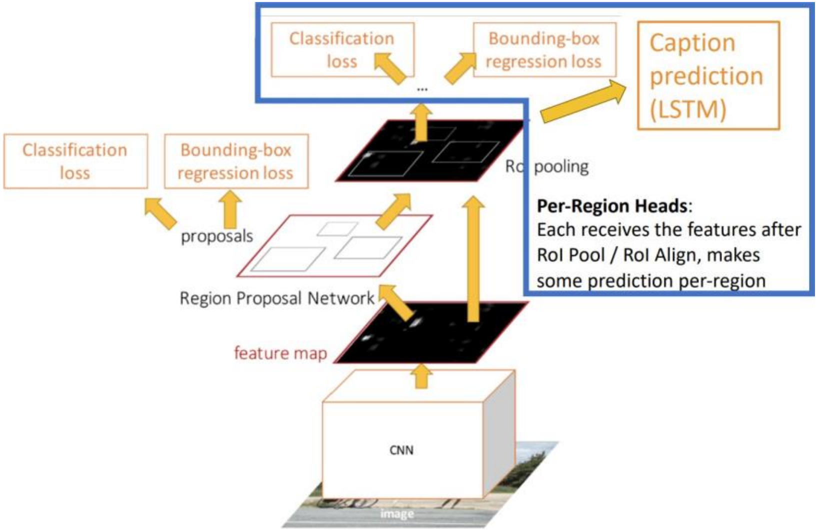
Mask R-CNN: Keypoint Estimation



Joint Instance Segmentation and Pose Estimation

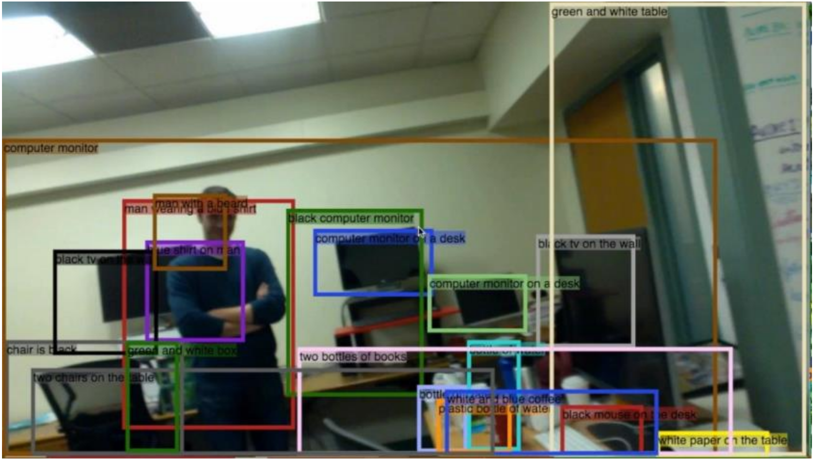
[0] He et al, “Mask R-CNN”, ICCV 2017

Captioning: Predict a caption per region!

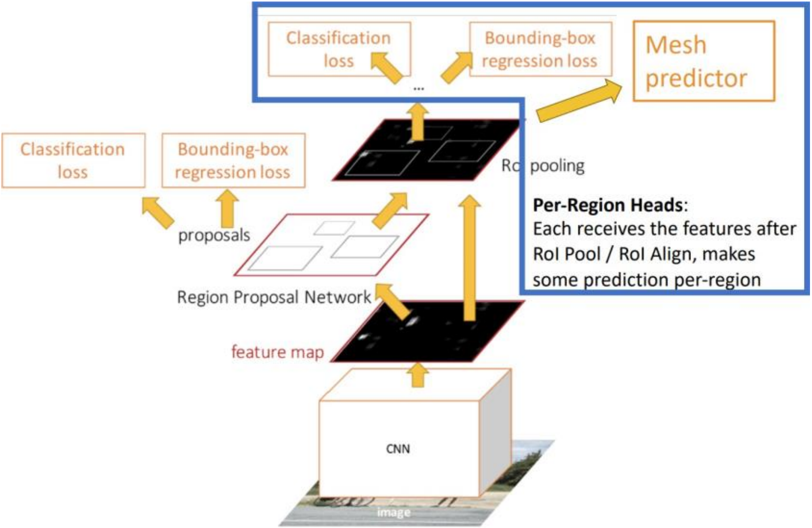


Per-Region Heads:
Each receives the features after RoI Pool / RoI Align, makes some prediction per-region

Captioning: Predict a caption per region!



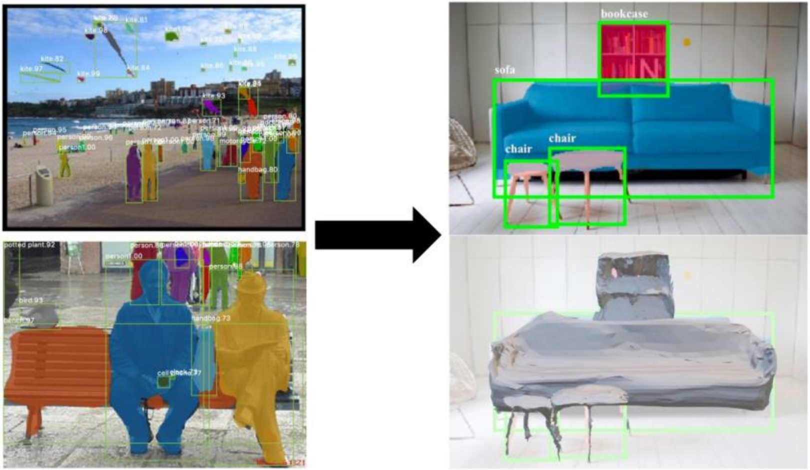
Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016


Classification loss
Bounding-box regression loss
Mesh predictor
...
Classification loss
Bounding-box regression loss
RoI pooling
Per-Region Heads:
Each receives the features after RoI Pool / RoI Align, makes some prediction per-region
proposals
Region Proposal Network
feature map
CNN
image

3D Shape Prediction

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

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



Gkioxari, Malik, and Johnson, ”Mesh R-CNN”, ICCV 2019

Object Tracking

- **Goal:** Track objects over a sequence of photos or a video
- Exceedingly challenging in multi-object tracking scenarios
- Need to take care of not mixing up or losing objects midway
- **One Solution:** Perform object detection and assign IDs to each object and store its feature vector. Then track the objects based on its ID and feature vector

Object Tracking

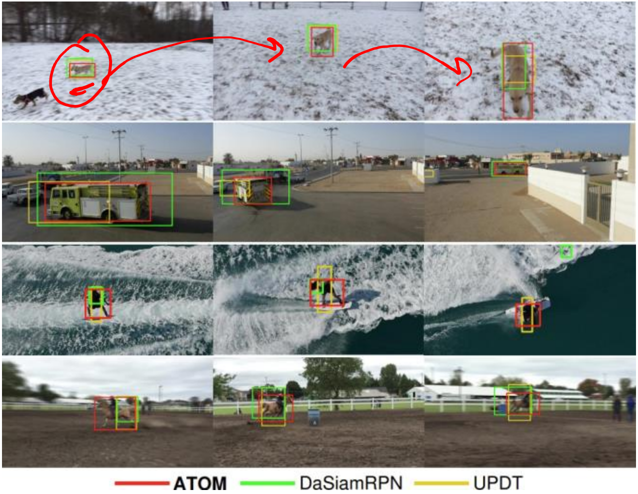


Figure 2: Comparison of 3 approaches for object tracking

Danelljan et al.

References

These slides have been adapted from

- Fei-Fei Li, Yunzhu Li & Ruohan Gao, Stanford CS231n: [Deep Learning for Computer Vision]
- Assaf Shocher, Shai Bagon, Meirav Galun & Tali Dekel, WAIC DL4CV [Deep Learning for Computer Vision: Fundamentals and Applications]
- Justin Johnson, UMich EECS 498.008/598.008: [Deep Learning for Computer Vision]