Deep Learning

Lecture 14

Object detection and segmentation

Part 2

Recap

Semantic Segmentation

GRASS , CAT ,
TREE , SKY

No objects, just pixels

2D Object Detection

DOG , DOG , CAT

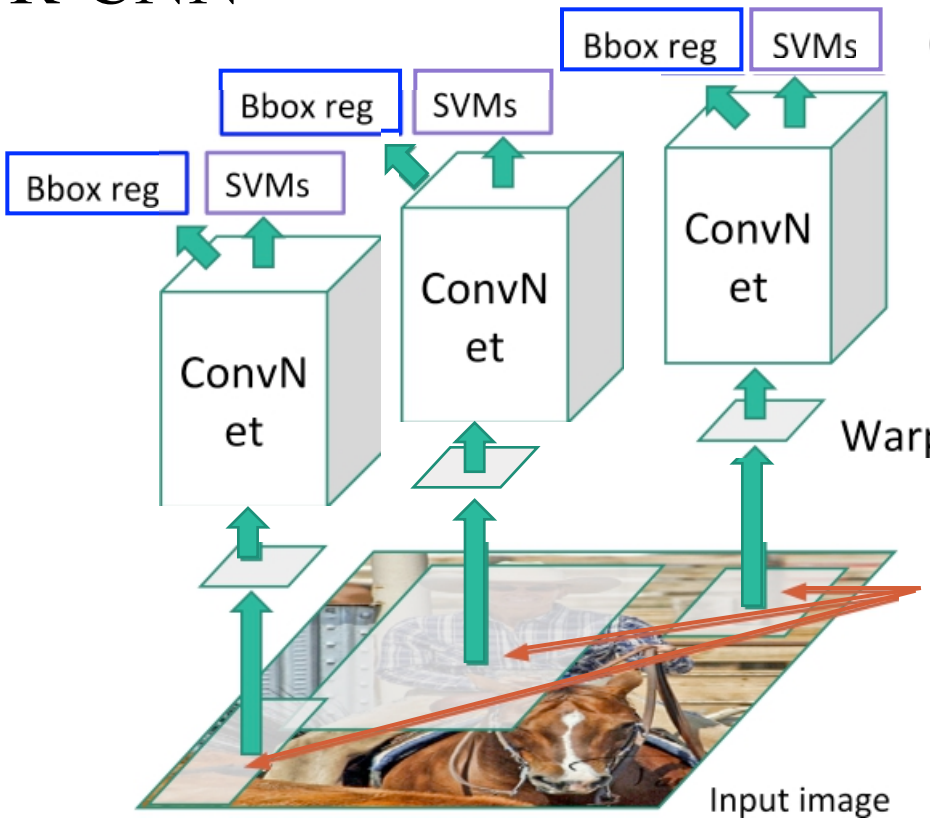
Object categories +
2D bounding boxes

3D Object Detection

Car

Object categories +
3D bounding boxes

R-CNN



Linear Regression for bounding box offsets

Classify regions with SVMs

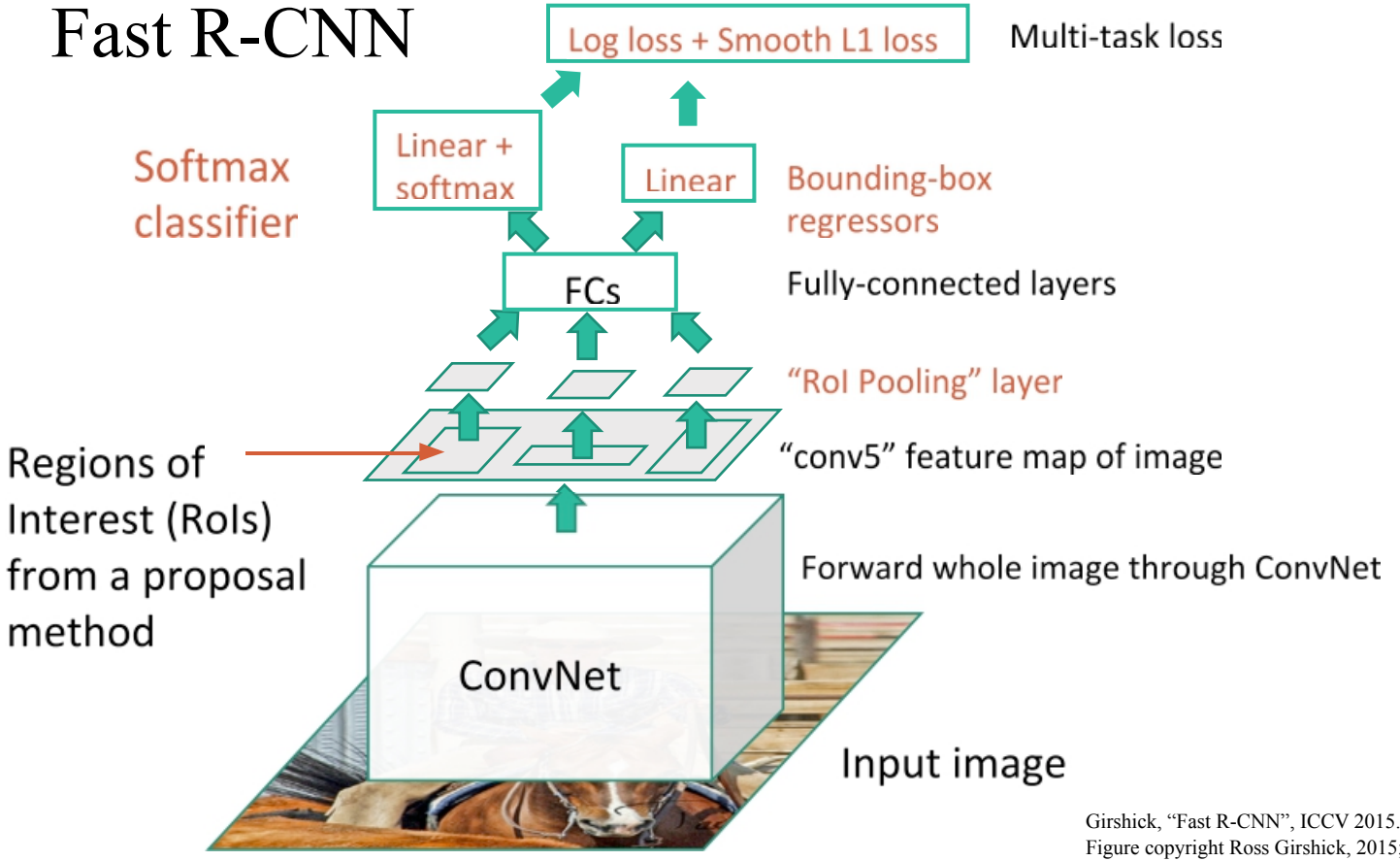
Forward each region through ConvNet

Warped image regions

Regions of Interest (RoI) from a proposal method (~2k)

Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast R-CNN



Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

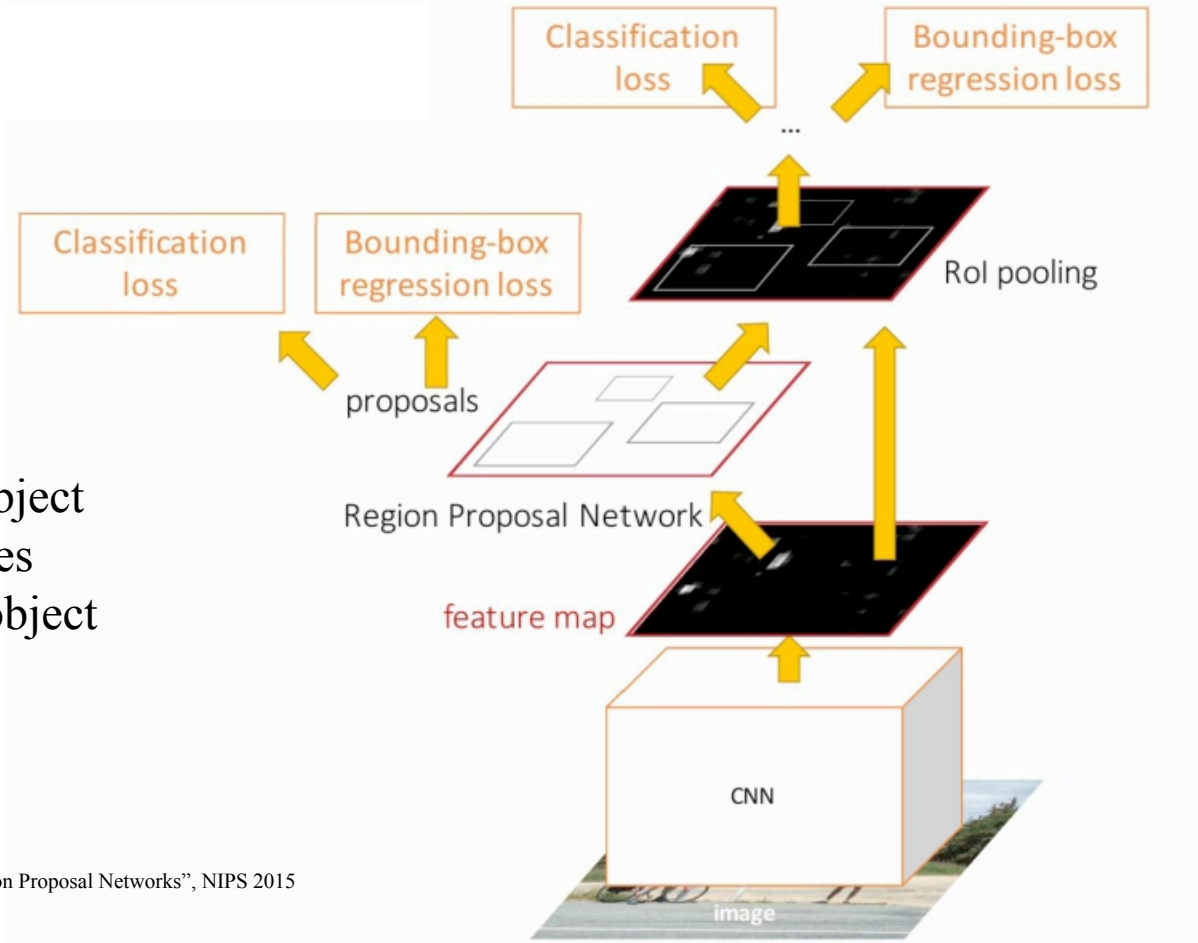
Faster R-CNN:

Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Jointly train with 4 losses:

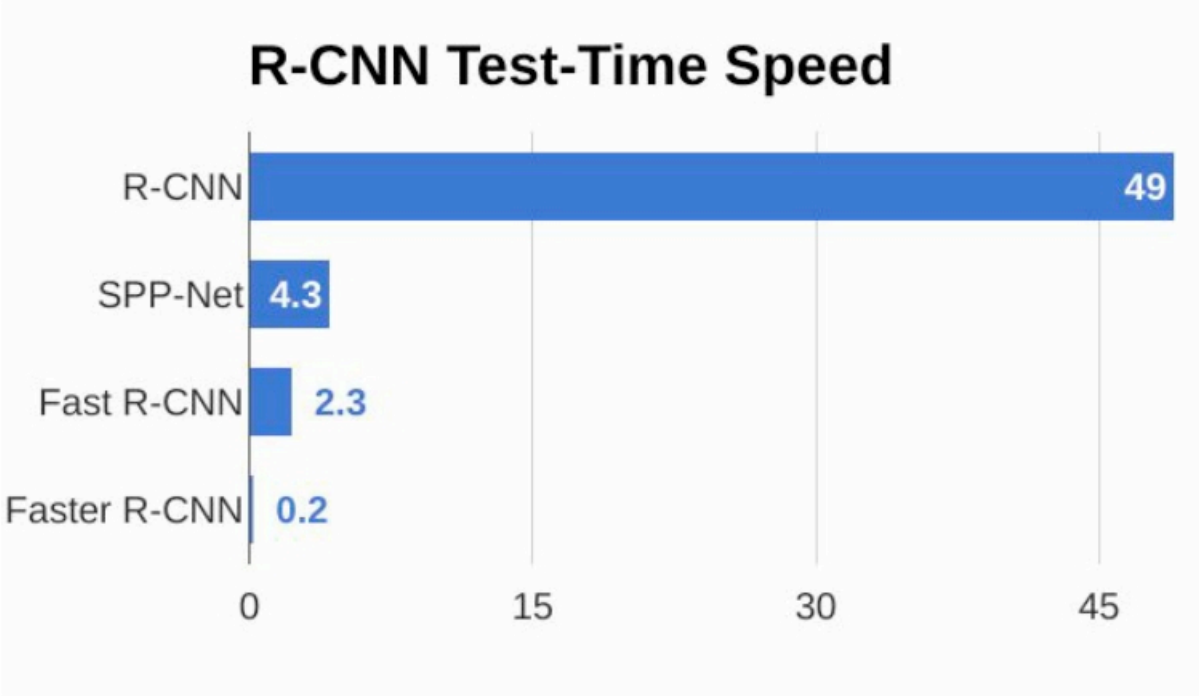
1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

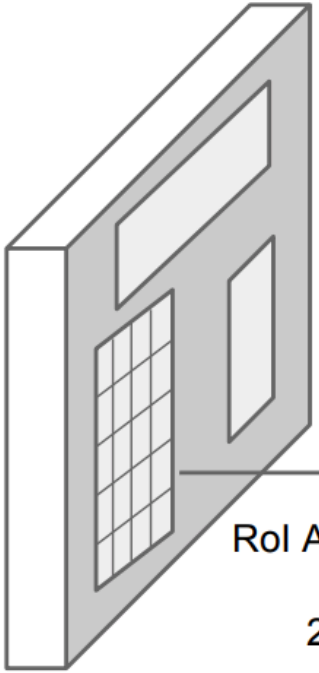
Fast **er** R-CNN:

Make CNN do proposals!



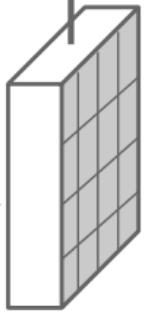
Mask R-CNN

CNN



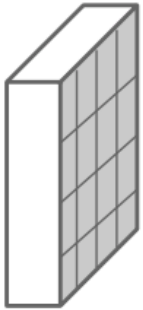
RoI Align

Classification Scores: C
Box coordinates (per class): 4 * C



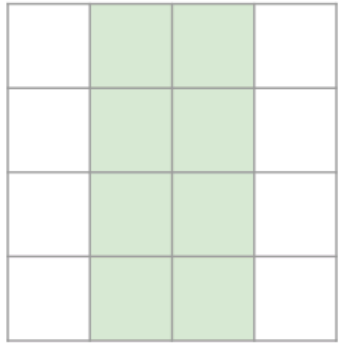
256 x 14 x 14

Conv



256 x 14 x 14

Conv

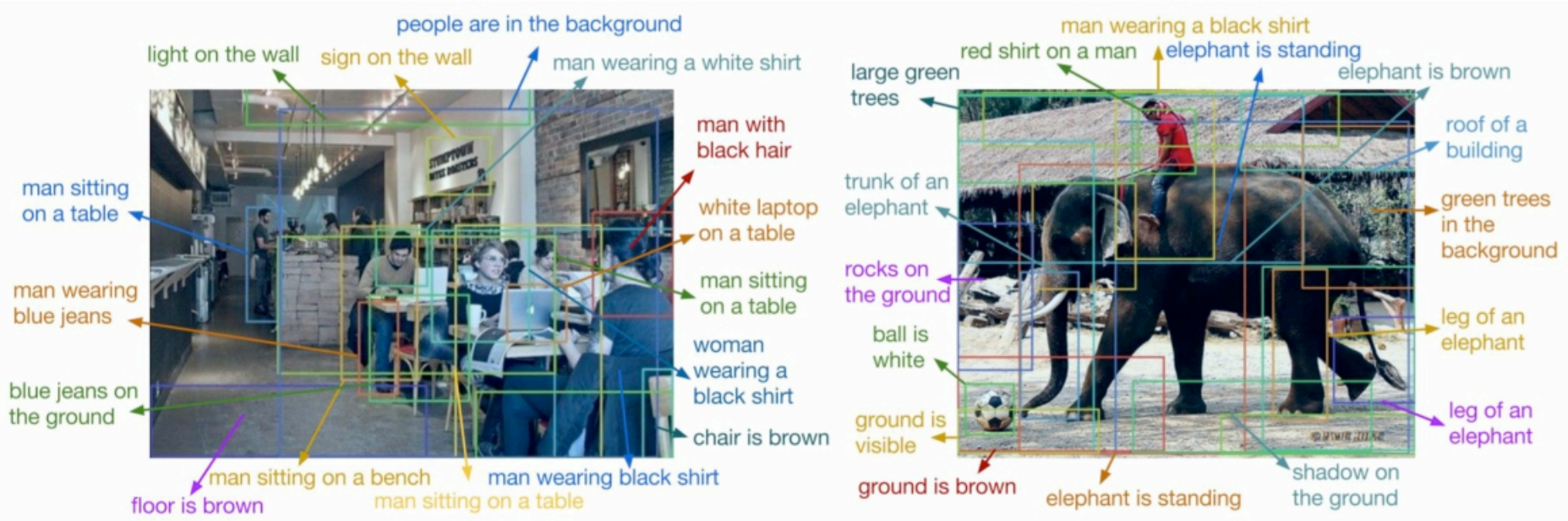


Predict a mask for each of C classes

C x 14 x 14

He et al, “Mask R-CNN”, arXiv 2017

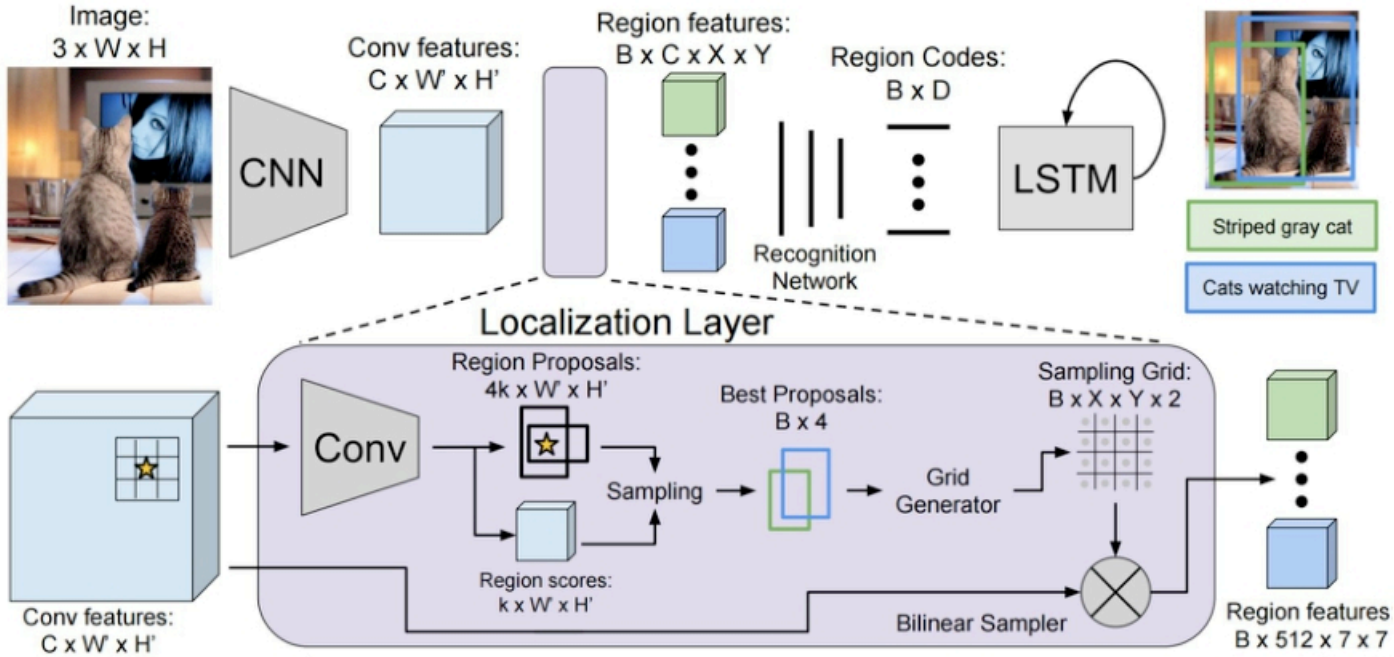
Object Detection + Captioning = Dense Captioning



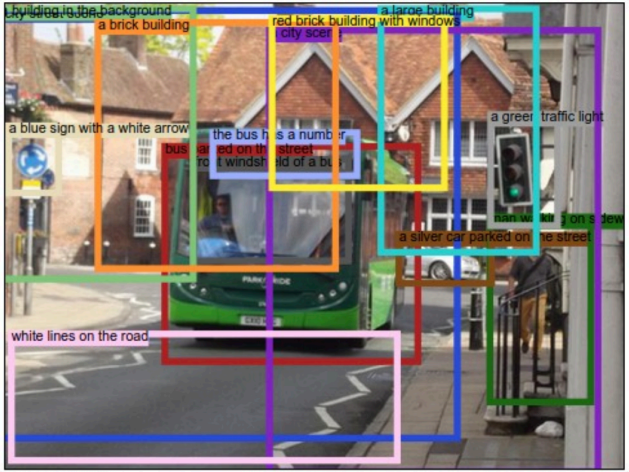
Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016
Figure copyright IEEE, 2016. Reproduced for educational purposes.

Used with permission from Justin Johnson

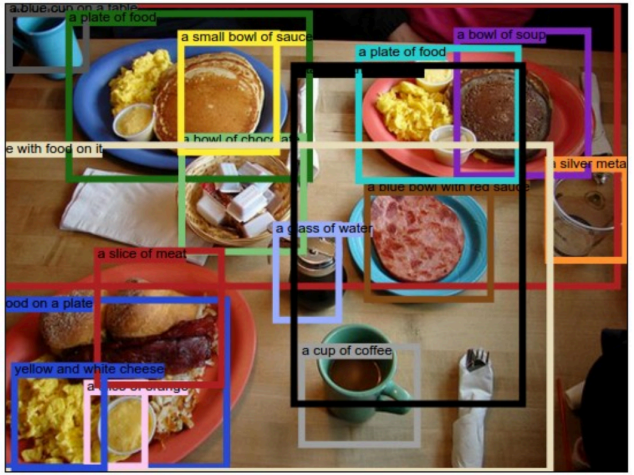
Object Detection + Captioning = Dense Captioning



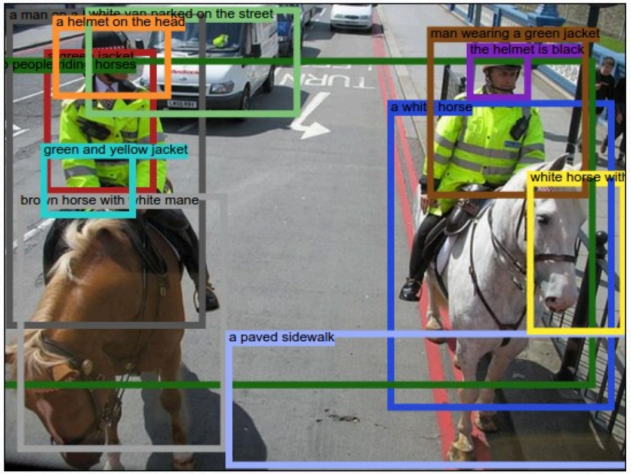
Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016
Figure copyright IEEE, 2016. Reproduced for educational purposes.



bus parked on the street. a city street scene. front windshield of a bus. man walking on sidewalk. a silver car parked on the street. a city scene. a green traffic light. a building in the background. the bus has a number. a large building. a brick building. red brick building with windows. a blue sign with a white arrow. white lines on the road.



a plate of food. food on a plate. a blue cup on a table. a plate of food. a blue bowl with red sauce. a bowl of soup. a cup of coffee. a bowl of chocolate. a glass of water. a plate of food. a silver metal container. a small bowl of sauce. table with food on it. a slice of orange. a table with food on it. a slice of meat. yellow and white cheese.



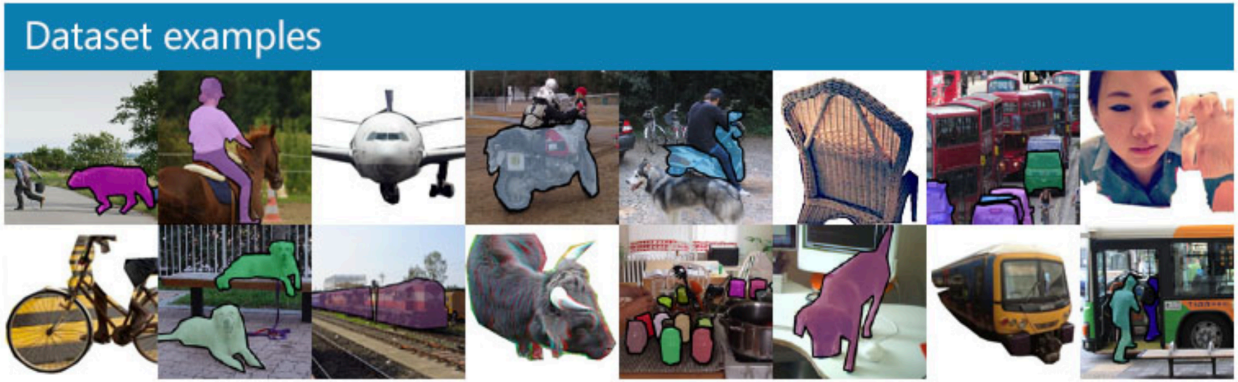
a green jacket. a white horse. a man on a horse. two people riding horses. man wearing a green jacket. the helmet is black. brown horse with white mane. white van parked on the street. a paved sidewalk. green and yellow jacket. a helmet on the head. white horse with white face.

Datasets: MS COCO

What is COCO?

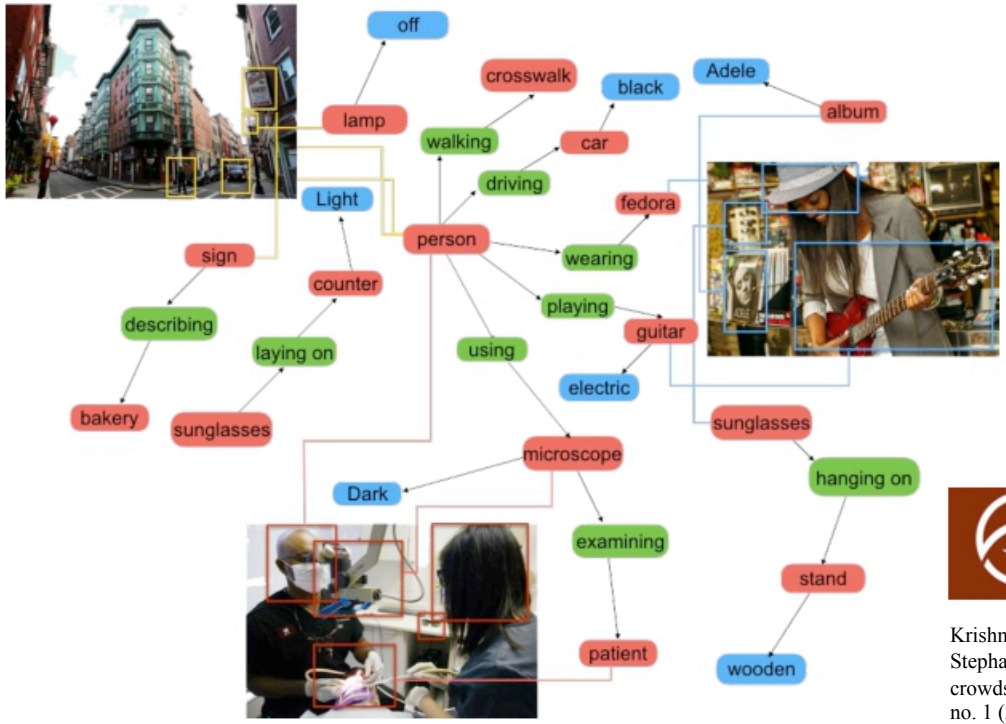
COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- **Object segmentation**
- **Recognition in context**
- **Superpixel stuff segmentation**
- **330K images (>200K labeled)**
- **1.5 million object instances**
- **80 object categories**
- **91 stuff categories**
- **5 captions per image**
- **250,000 people with keypoints**



Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755)

Datasets: VisualGenome



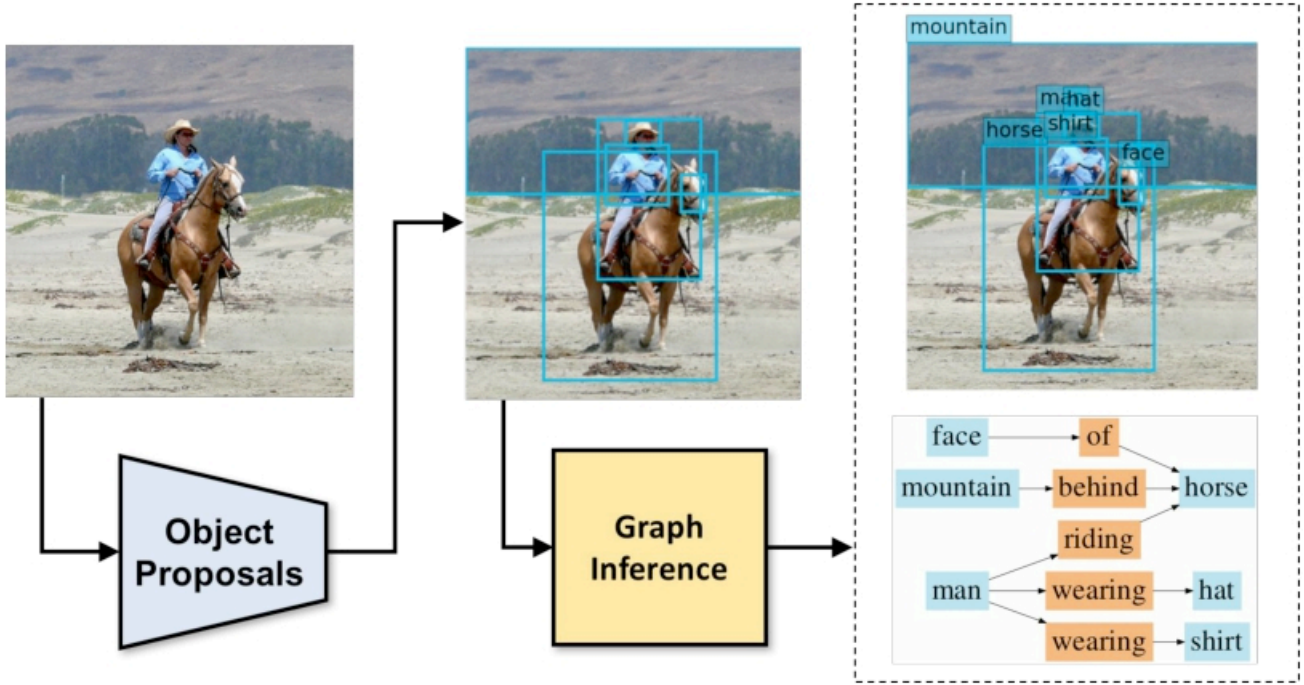
108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answers
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Wordnet Synsets

VISUALGENOME

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123, no. 1 (2017): 32-73.

Used with permission from Justin Johnson

Aside: Scene Graph Generation



Xu, Zhu, Choy, and Fei-Fei, "Scene Graph Generation by Iterative Message Passing", CVPR 2017
Figure copyright IEEE, 2018. Reproduced for educational purposes.

3D Object Detection

3D Object Detection

Car

Object categories +
3D bounding boxes

Simplified Camera Model

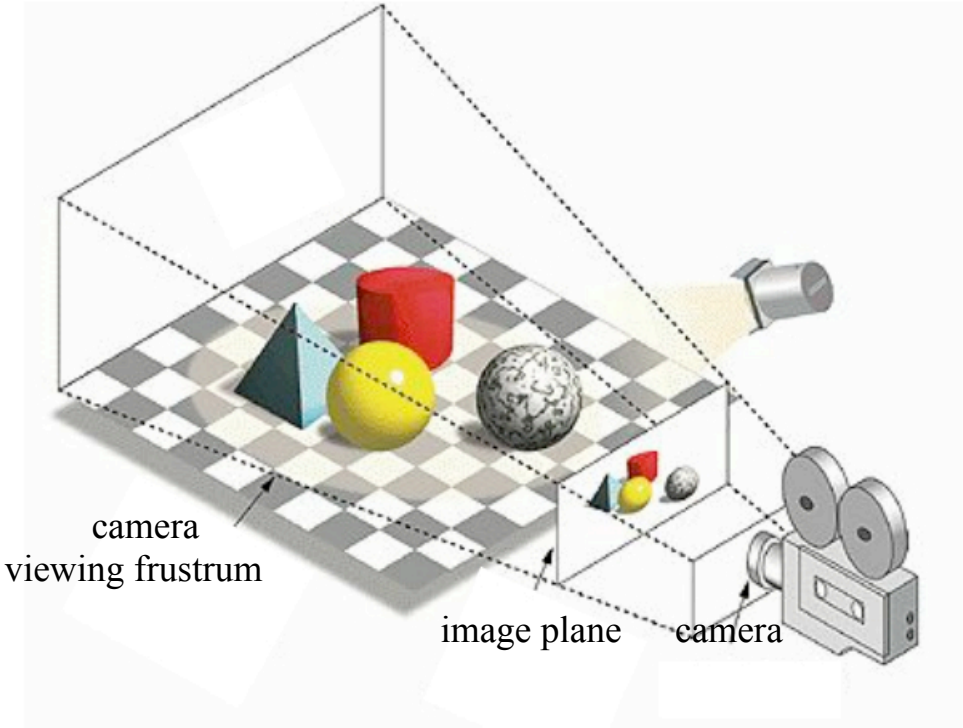
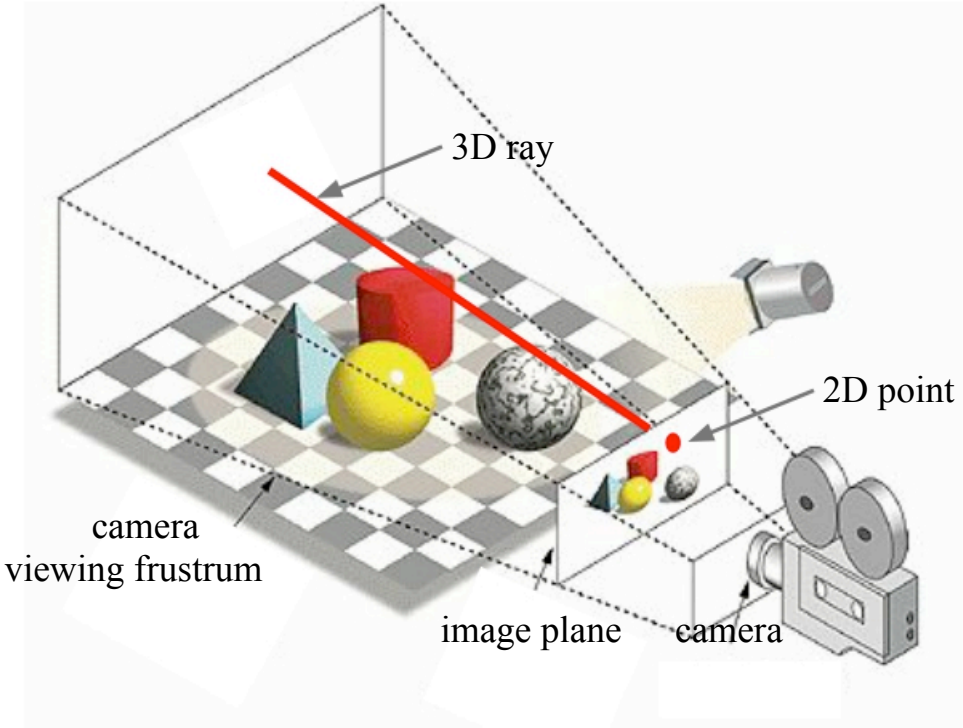


Image source: https://www.pcmag.com/encyclopedia_images/_FRUSTUM.GIF

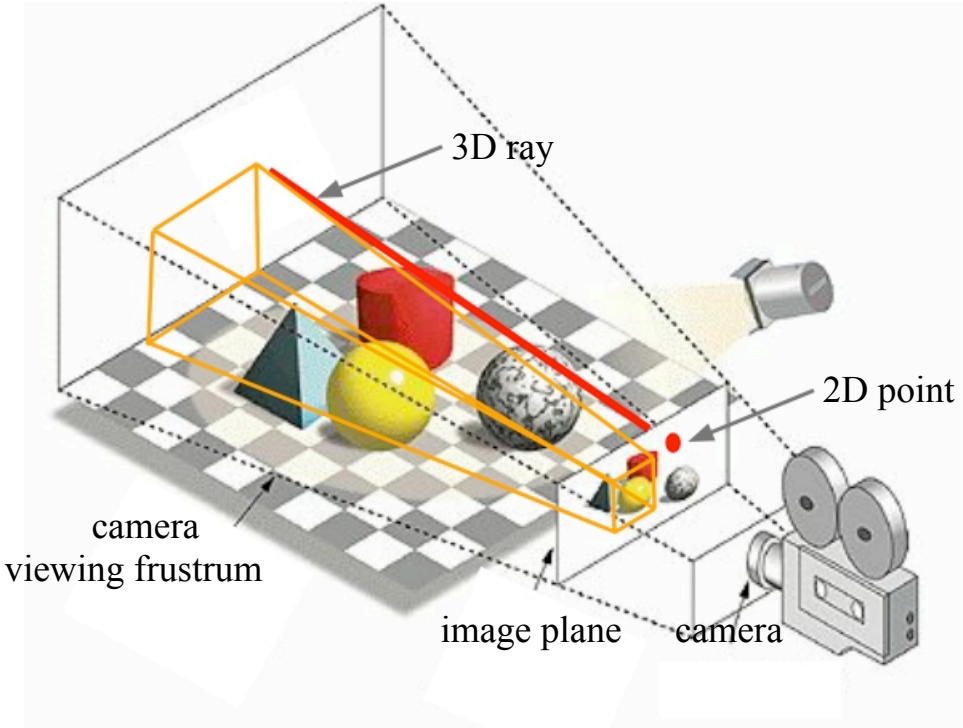
Simplified Camera Model



A point on the image plane corresponds to a **ray** in the 3D space

Image source: https://www.pcmag.com/encyclopedia_images/_FRUSTUM.GIF

Simplified Camera Model

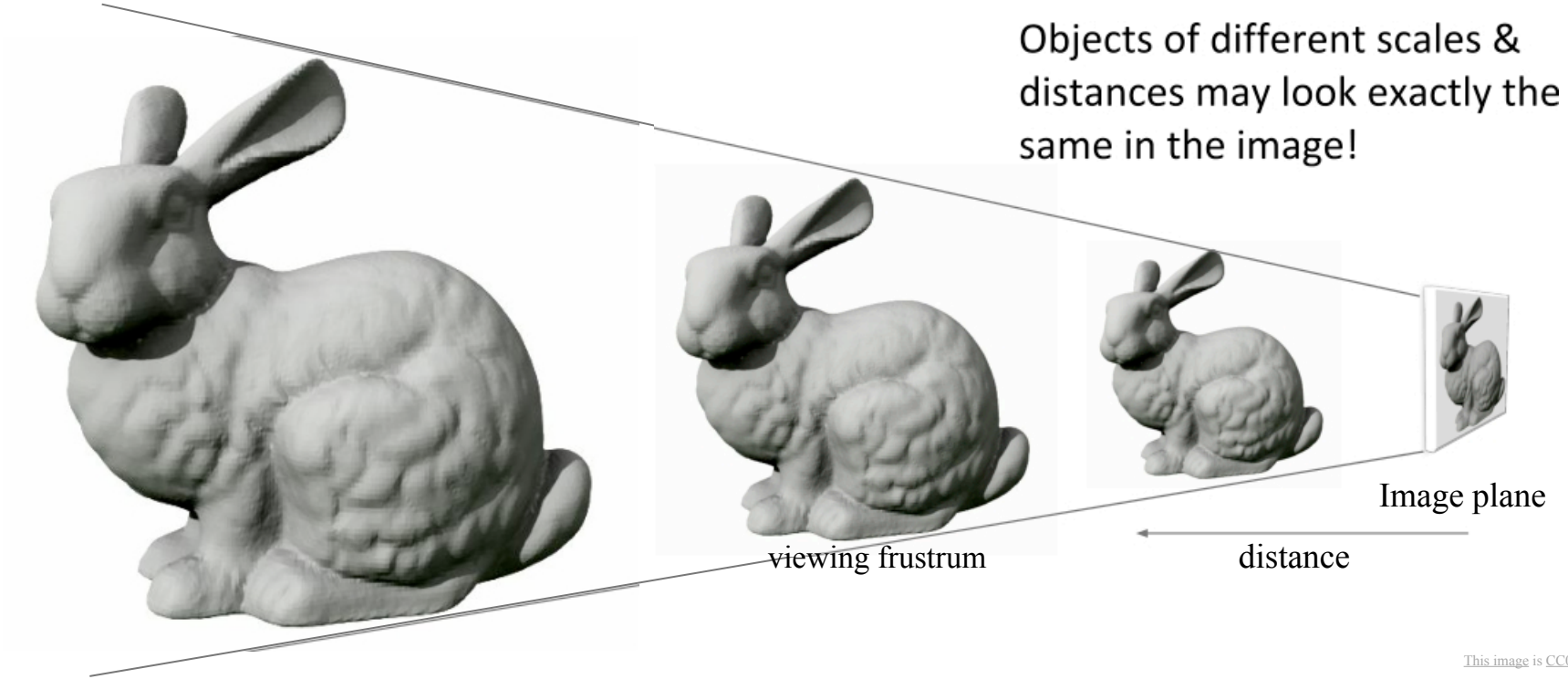


A point on the image plane corresponds to a **ray** in the 3D space

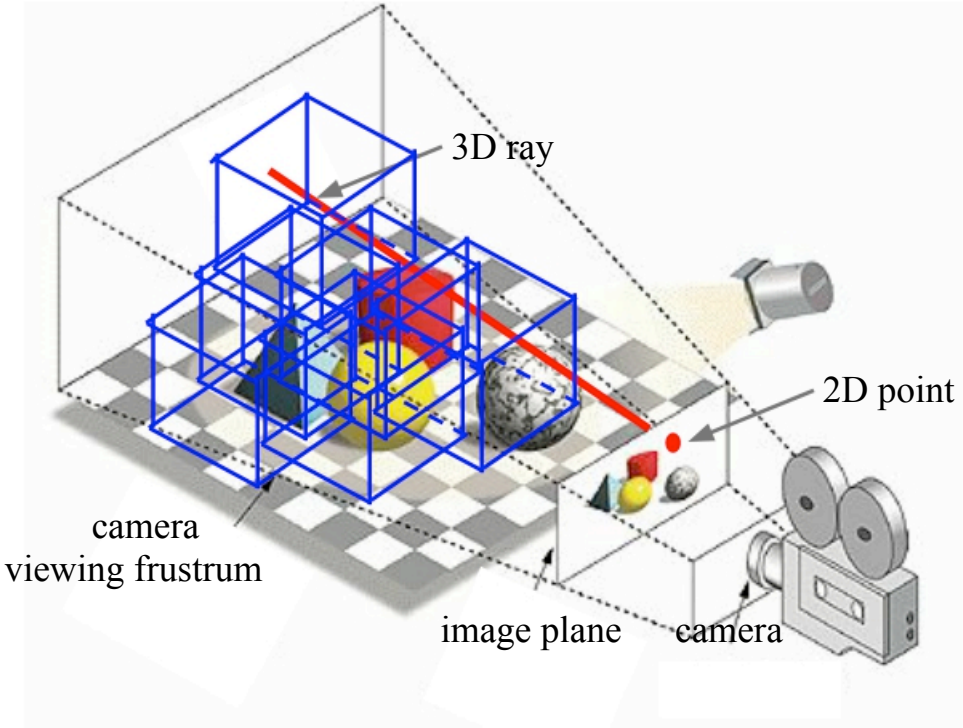
A 2D bounding box on an image is a **frustrum** in the 3D space

Image source: https://www.pcmag.com/encyclopedia_images/_FRUSTUM.GIF

Scale & Distance Ambiguity



Simplified Camera Model



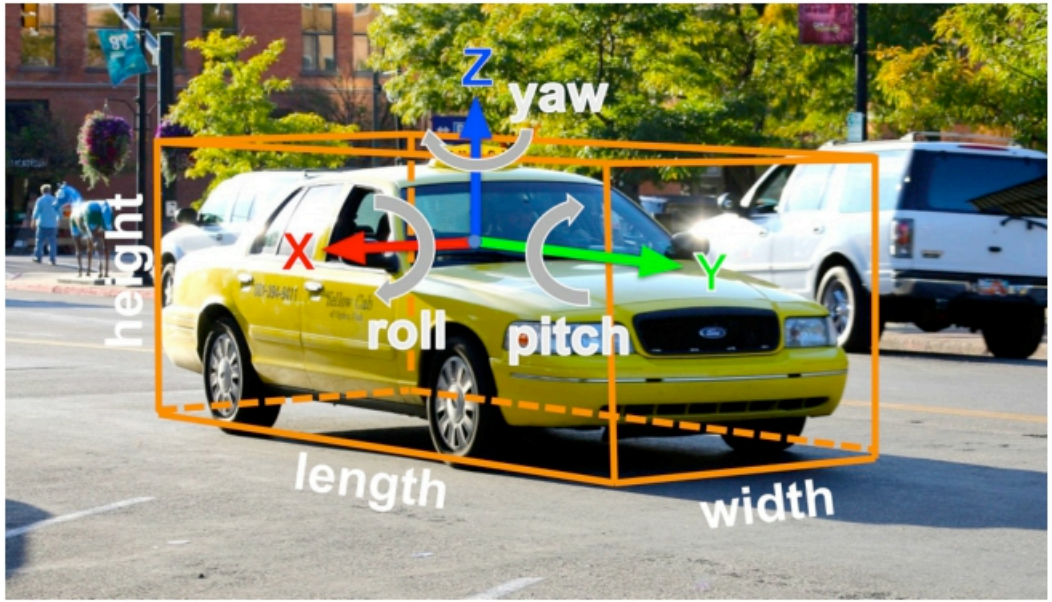
A point on the image plane corresponds to a **ray** in the 3D space

A 2D bounding box on an image is a **frustrum** in the 3D space

Localize an object in 3D:
The object can be anywhere in the **camera viewing frustrum**!

Image source: https://www.pcmag.com/encyclopedia_images/_FRUSTUM.GIF

3D Object Detection



2D Object Detection:
2D bounding box
(x, y, w, h)

3D Object Detection:
3D oriented bounding box
(x, y, z, w, h, l, r, p, y)

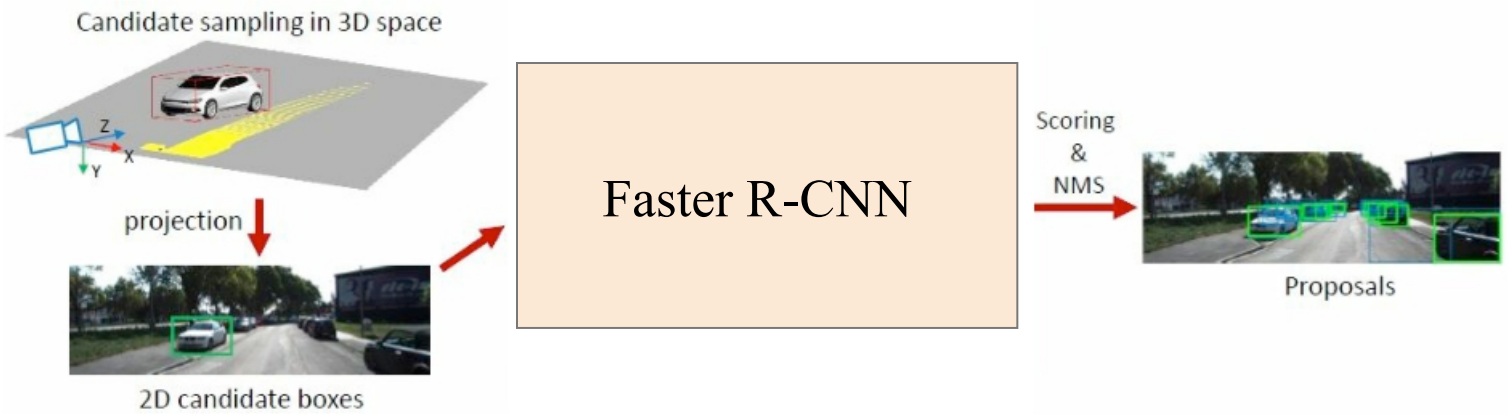
Simplified bbox: no roll & pitch

Much harder problem than 2D object detection!

3D Object Detection

Used with permission from Justin Johnson

3D Object Detection: Monocular Camera

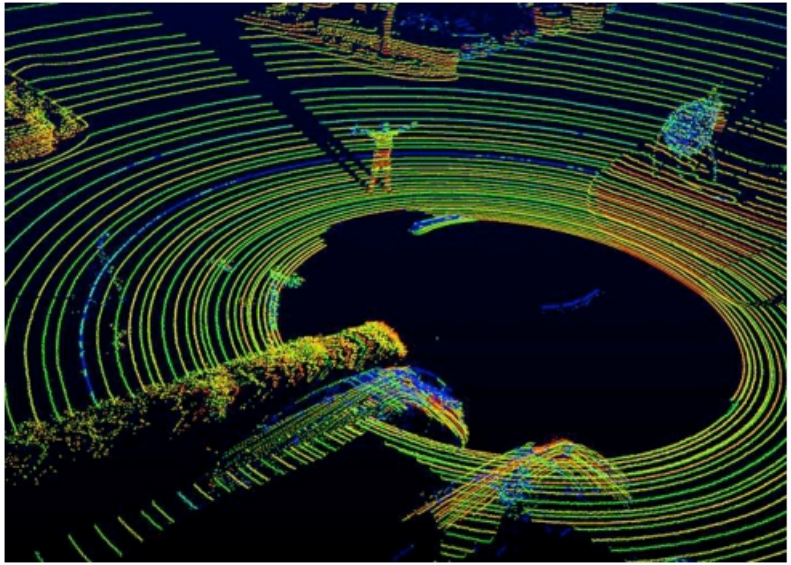


- Same idea as Faster RCNN, but proposals are in 3D
- 3D bounding box proposal, regress 3D box parameters + class score

Chen, Xiaozhi, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. "Monocular 3d object detection for autonomous driving." CVPR 2016.

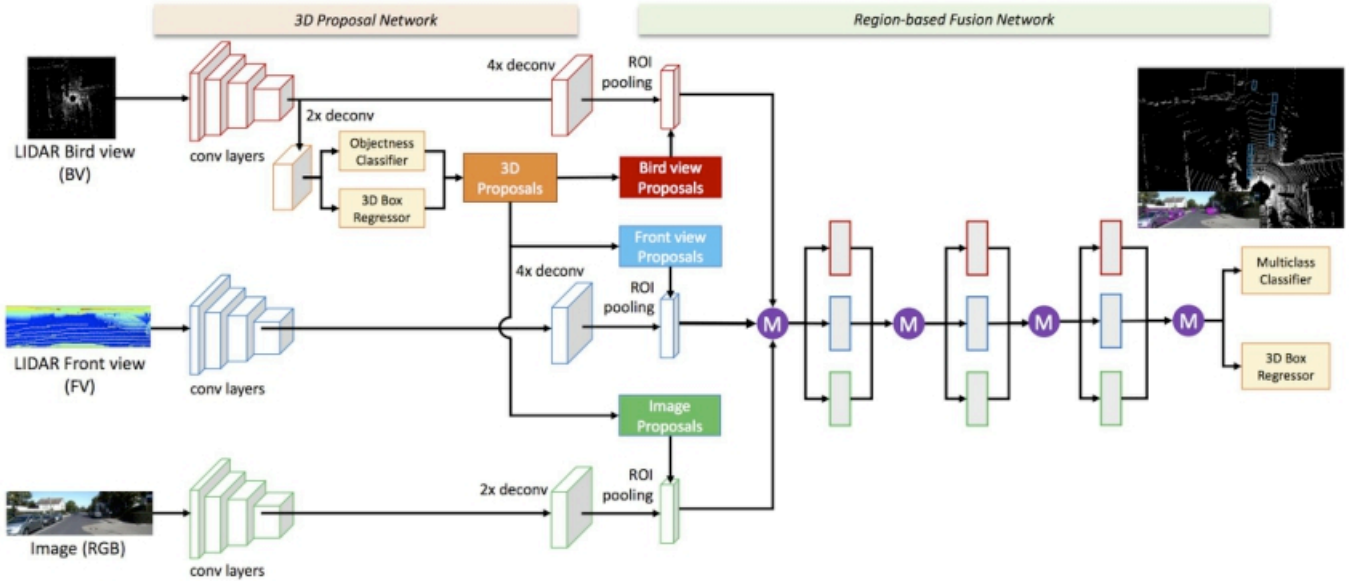
3D Object Detection: Camera + LiDAR

Velodyne (HDL-64e)



3D Point Cloud

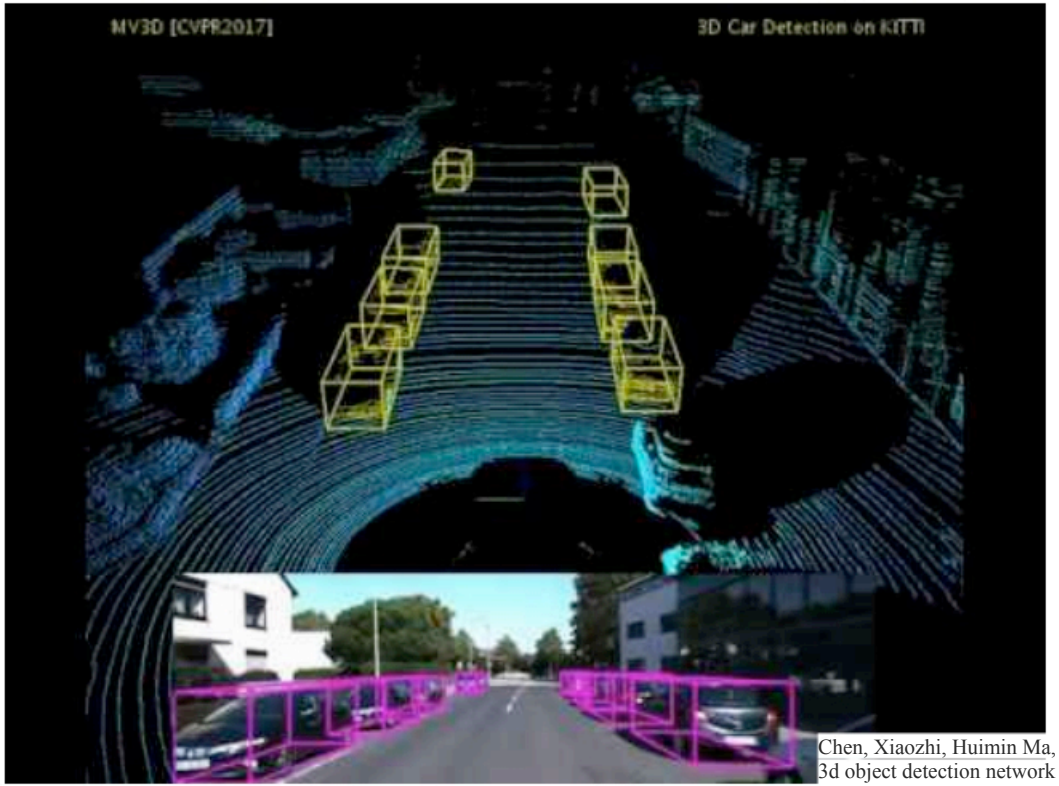
3D Object Detection: Camera + LiDAR



- Combine 3D proposals from multiple views & sensors
- regress 3D box parameters + class score

Chen, Xiaozhi, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. "Multi-view 3d object detection network for autonomous driving." *CVPR* 2017

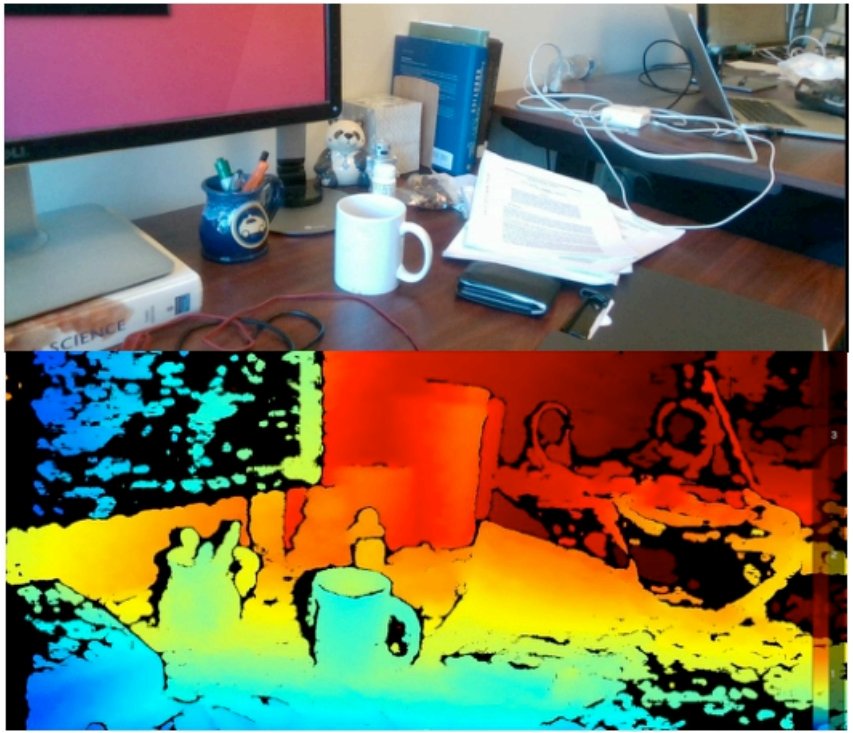
3D Object Detection: Camera + LiDAR



Chen, Xiaozhi, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. "Multi-view 3d object detection network for autonomous driving." *CVPR* 2017

RGB-Depth Camera

Kinect (Xbox One)



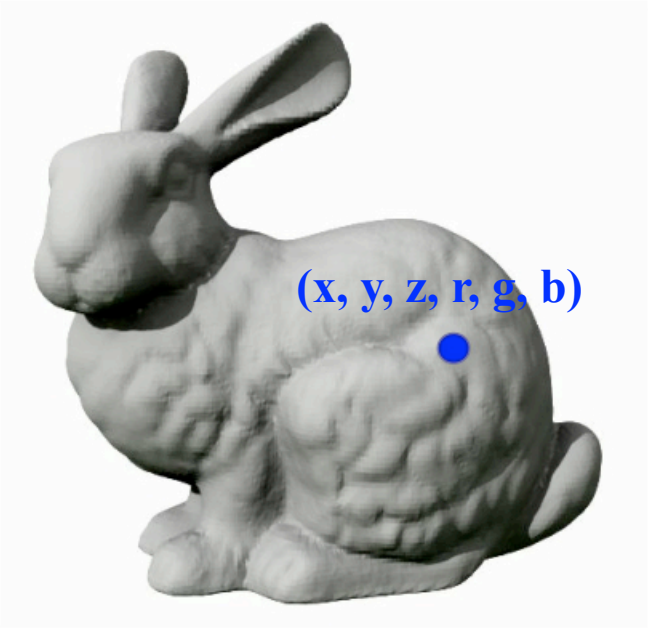
RGB

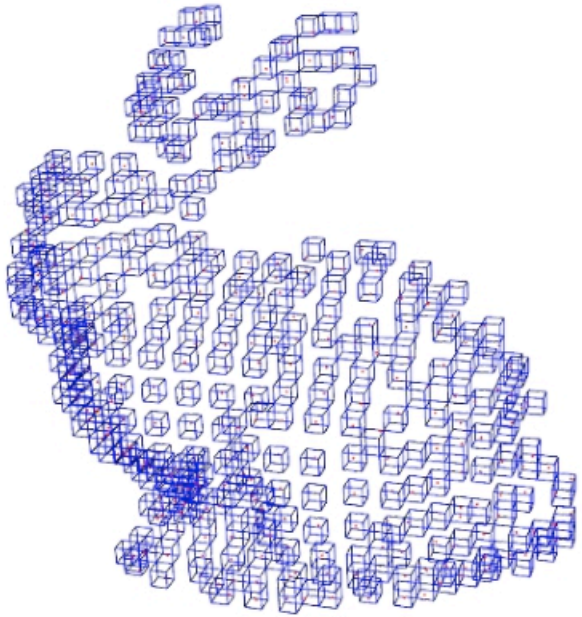
Depth

RGB-Depth Camera

Registered RGB + depth point cloud

Point Cloud Voxelization





1. Capture RGB-D point cloud of a scene.

2. Partition the 3D space into a regular 3D grid.

3. For each grid cell that has a point fall into it, fill the cell with the RGB value of that point.

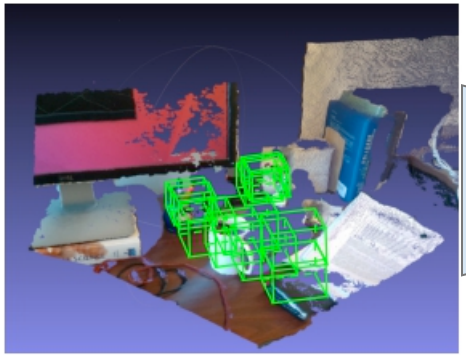
A bit like “3D image”

3D Object Detection: RGB-Depth Camera

Voxelized RGB-D point cloud

3D Conv

R P N



3D region proposals

More layers

3D NMS

Object categories + 3D bounding boxes

"Faster RCNN in 3D"

S. Song, and J. Xiao. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. CVPR 2016

Recap

Semantic Segmentation

GRASS , CAT ,
TREE , SKY

No objects, just pixels

2D Object Detection

DOG , DOG , CAT

Object categories +
2D bounding boxes

3D Object Detection

Car

Object categories +
3D bounding boxes