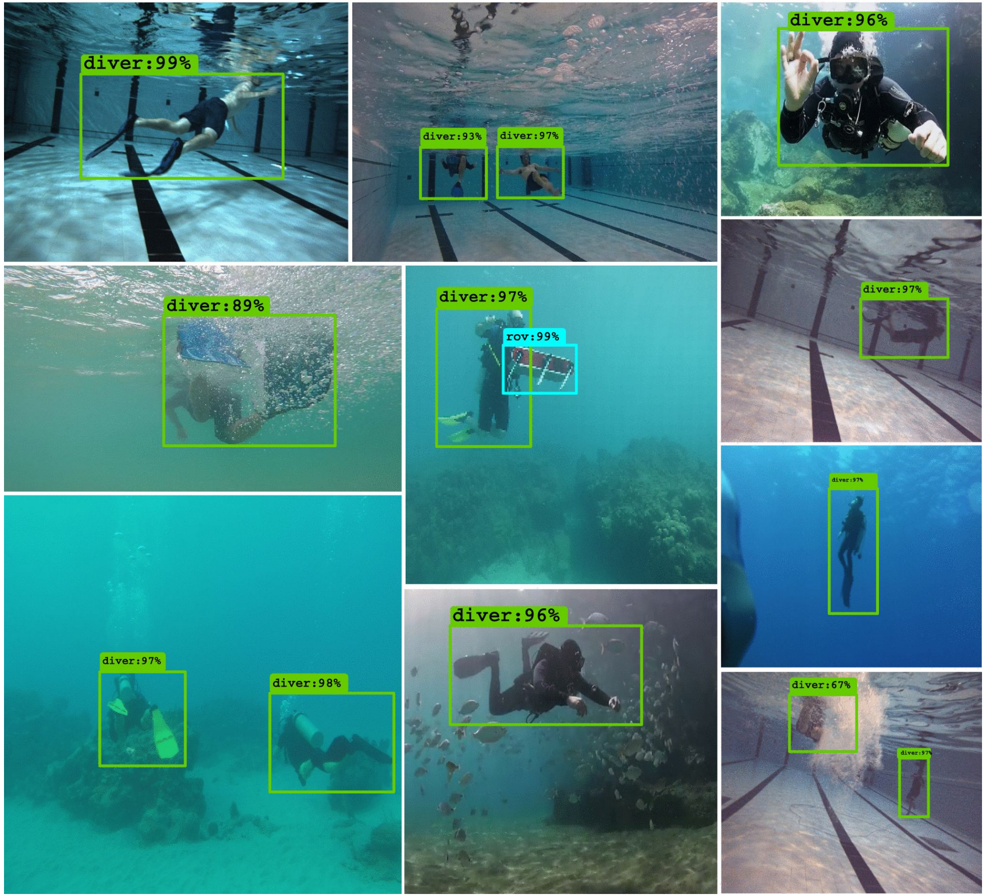
CSCI 8980 Think Deep Learning

Applications of CNNs in Computer Vision

Applications of CNNs in Computer Vision

- Object Detection
- Segmentation

Object Detection



(Islam et al., "Toward a Generic Diver-Following Algorithm: Balancing Robustness and Efficiency in Deep Visual Detection," 2019)

Paper from UMN IRVLab: http://irvlab.dl.umn.edu/

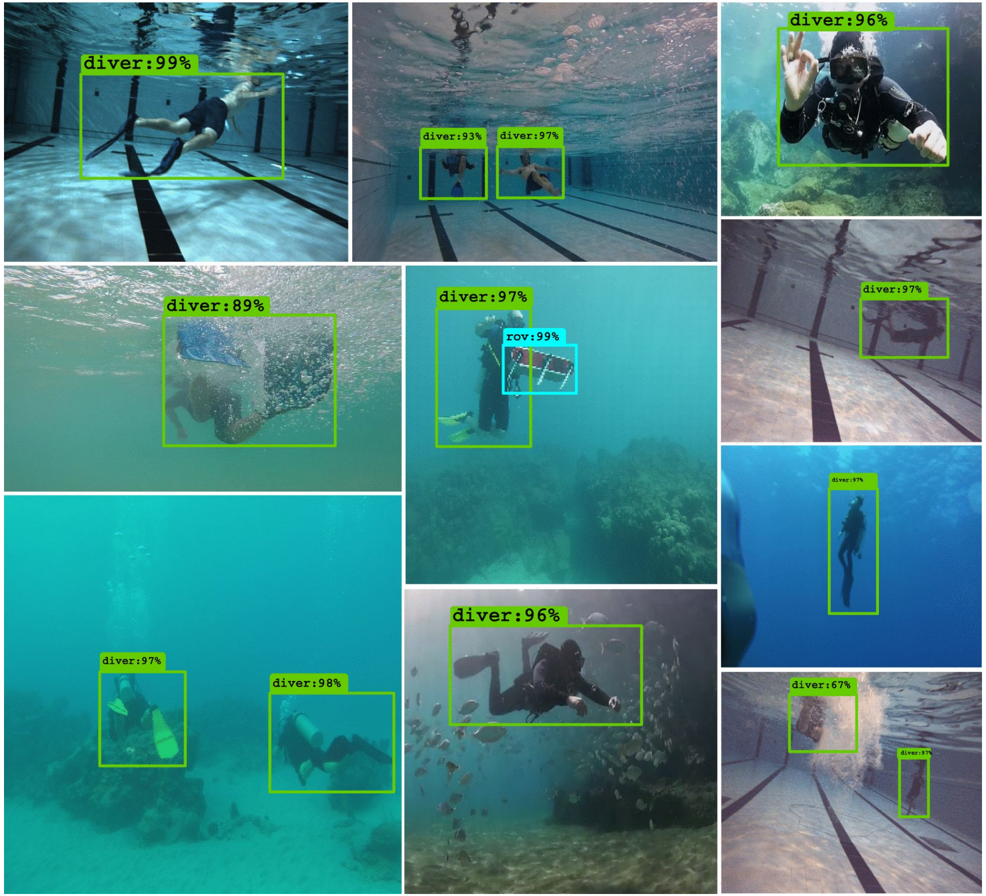
What is Object Detection?

2 main tasks:

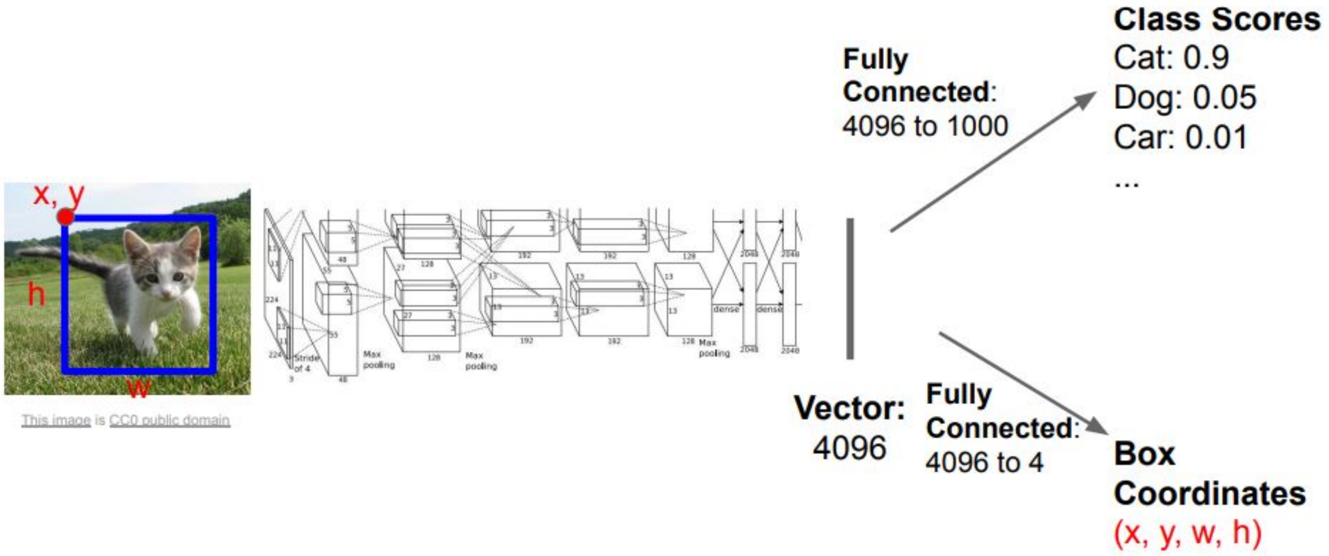
- **Localizing one or more objects in the Image**
- **Classifying each object in the image**

(Islam et al., "Toward a Generic Diver-Following Algorithm: Balancing Robustness and Efficiency in Deep Visual Detection," 2019)

Paper from UMN IRVLab: http://irvlab.dl.umn.edu/

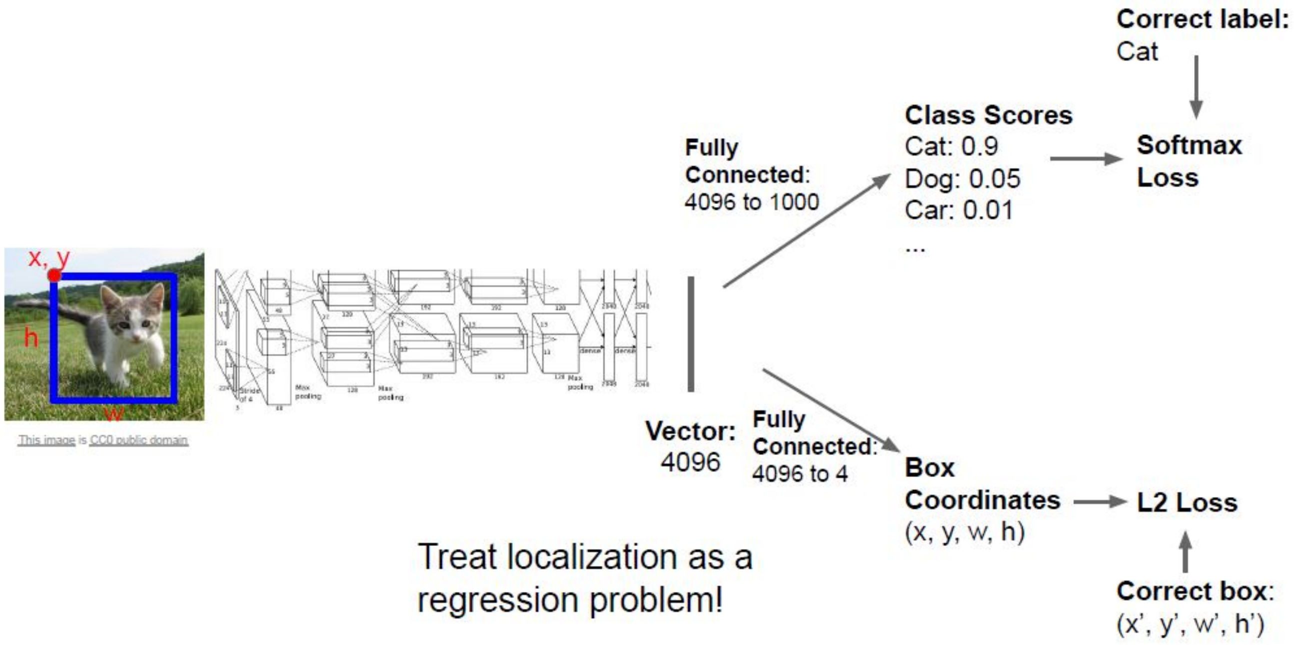


Object Detection Network Input and Output



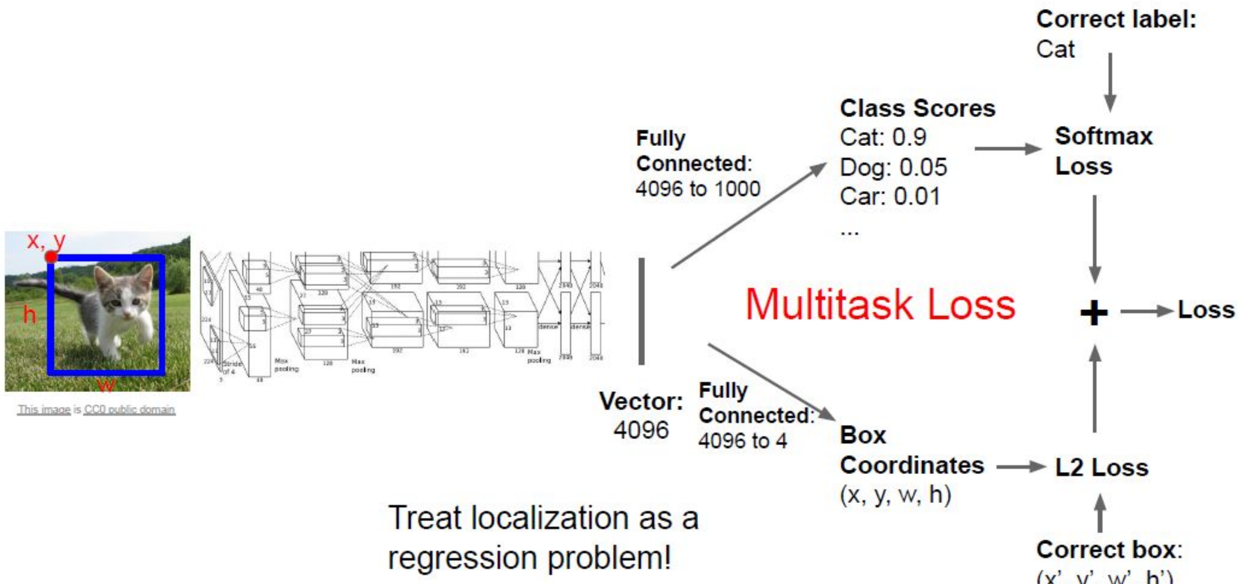
(Li et al., *Detection and Segmentation* 2020)

Object Detection Network: Training



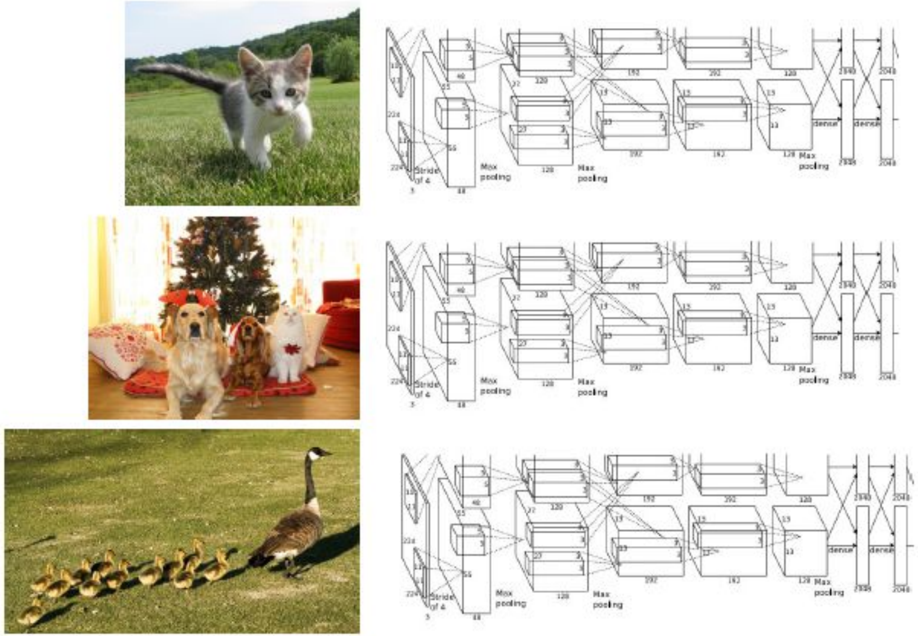
(Li et al., *Detection and Segmentation* 2020)

Object Detection Network: Training



(Li et al., *Detection and Segmentation* 2020)

Multiple Objects: Multiple outputs



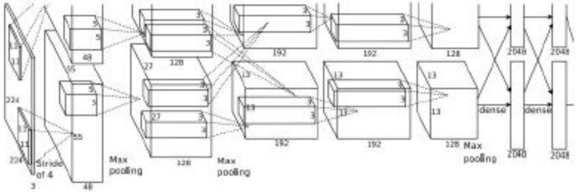
CAT: (x, y, w, h)

DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)

DUCK: (x, y, w, h)
DUCK: (x, y, w, h)
....

(Li et al., *Detection and Segmentation* 2020)

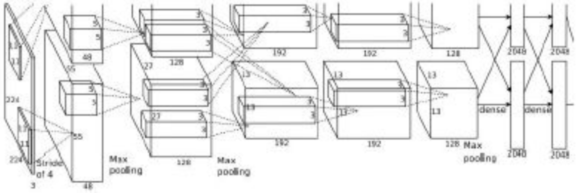
Multiple Objects: Initial Solution



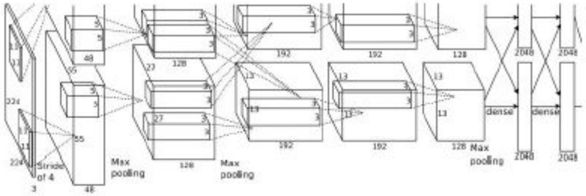
Dog? NO
Cat? NO
Background? YES

(Li et al., *Detection and Segmentation* 2020)

Multiple Objects: Initial Solution



Dog? NO
Cat? NO
Background? YES

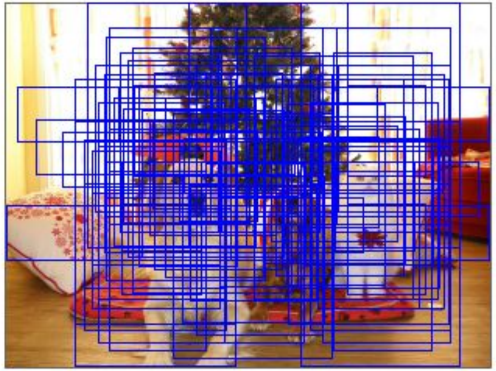


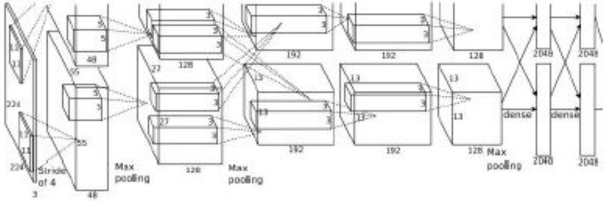
Dog? YES
Cat? NO
Background? NO

(Li et al., *Detection and Segmentation* 2020)

Multiple Objects: Heavy computational cost

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background





Dog? NO
Cat? YES
Background? NO

Problem: Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

(Li et al., *Detection and Segmentation* 2020)

Solution: 4-step Object-detection framework

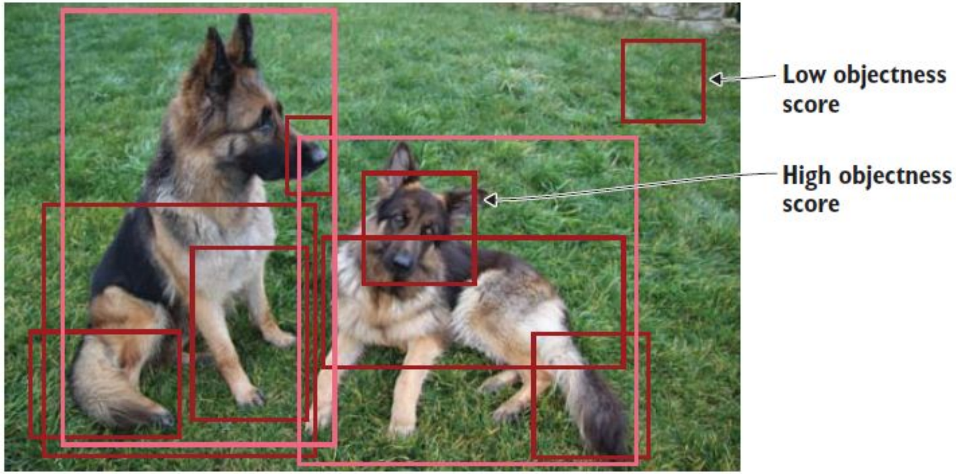
1. **Region Proposal:** identify regions of interest (RoI) for potential locations of objects
2. **Feature extraction:** extract visual features within each RoI for classification
3. **Non-maximum Suppression:** avoid repeated detections
4. **Evaluation metrics:** evaluate performance of model

1. Region Proposal:

Identification of Regions of Interest (RoIs)

- General procedures for Region Proposal:
  - Generate thousands of bounding boxes (BBs)
  - Classify BBs as foreground or background based on 'objectness score'
  - Pass only foreground through rest of network

- One common approach is using **Selective Search**
  - Fast algorithm, ~200 region proposals in A few seconds on CPU



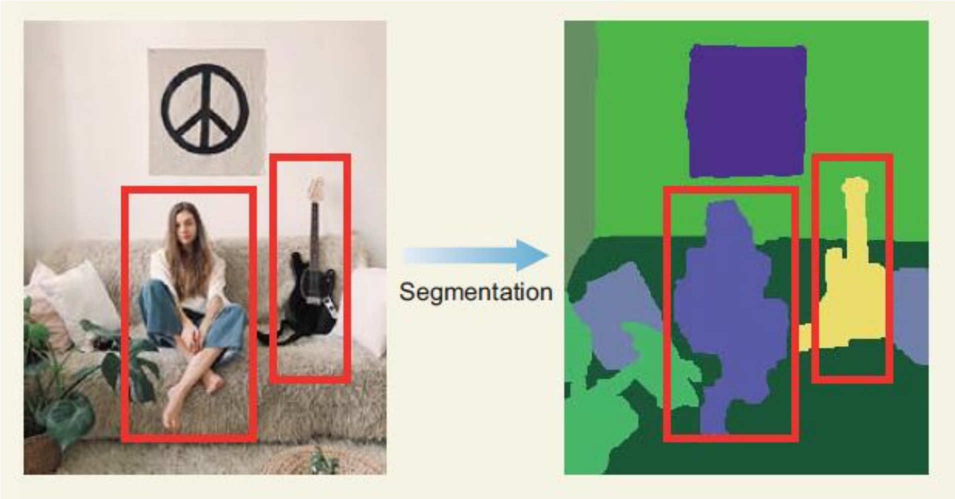
(Elgendy, 2020)

Selective Search

Greedy search algorithm for region proposal

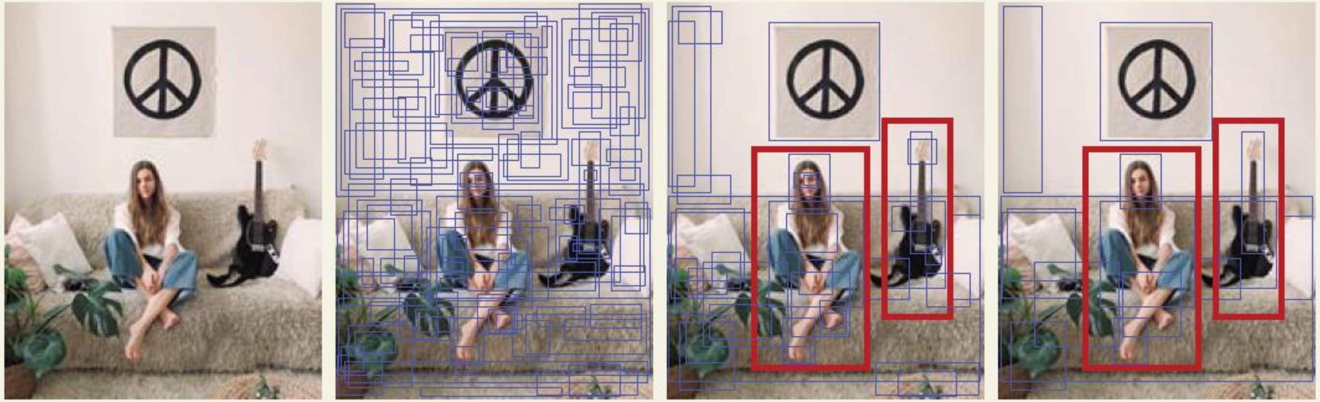
Step 1:

- Segmentation
- Defining ‘blobs’ that
  Could be objects



(Elgendy, 2020)

Selective Search



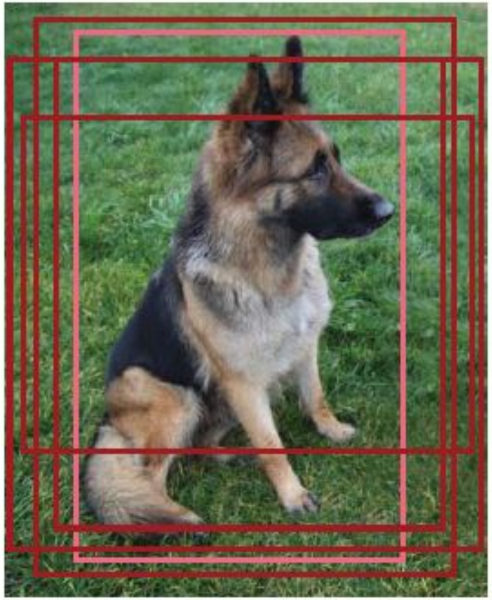
Input image | Proposed regions | After the first iteration | After a few iterations

(Elgendy, 2020)

2. Feature Extraction & Classification in RoI

Using a pretrained CNN network,

- Extract features using a pretrained CNN
- Make 2 predictions using additional layers:
  - Bounding box prediction (x, y, width, height)
  - Class prediction (softmax function predicting the class probability For each object

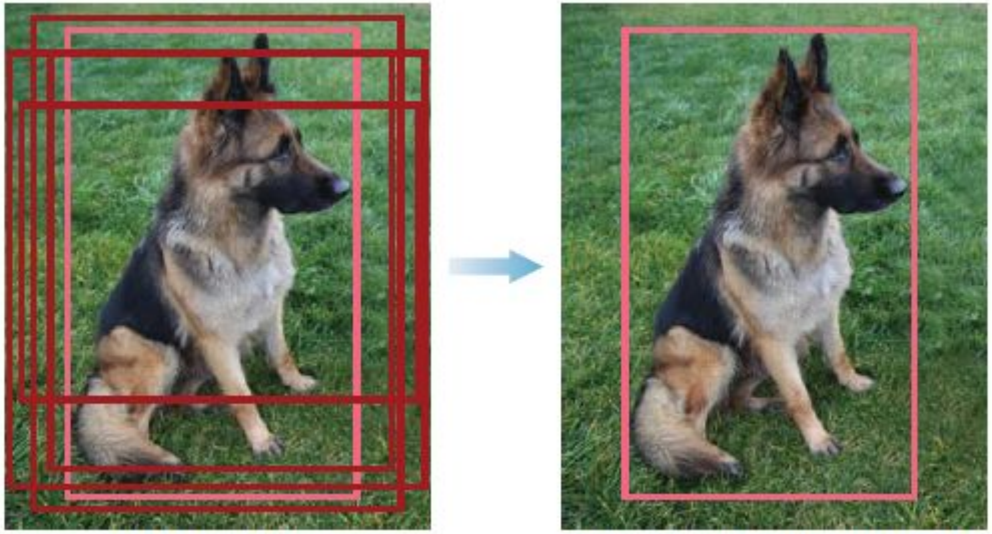


(Elgendy, 2020)

3. Remove Duplicate Object Detections

Non-maximum suppression (NMS):
A 4-step technique for eliminating duplicate detections of objects.

1. Discard BBs with predictions below a **confidence threshold**.
2. Select the BB with the highest probability
3. Calculate the overlap of all remaining boxes with the same class prediction
4. Suppress any box with an IoU smaller than a threshold (NMS threshold, usually 0.5).



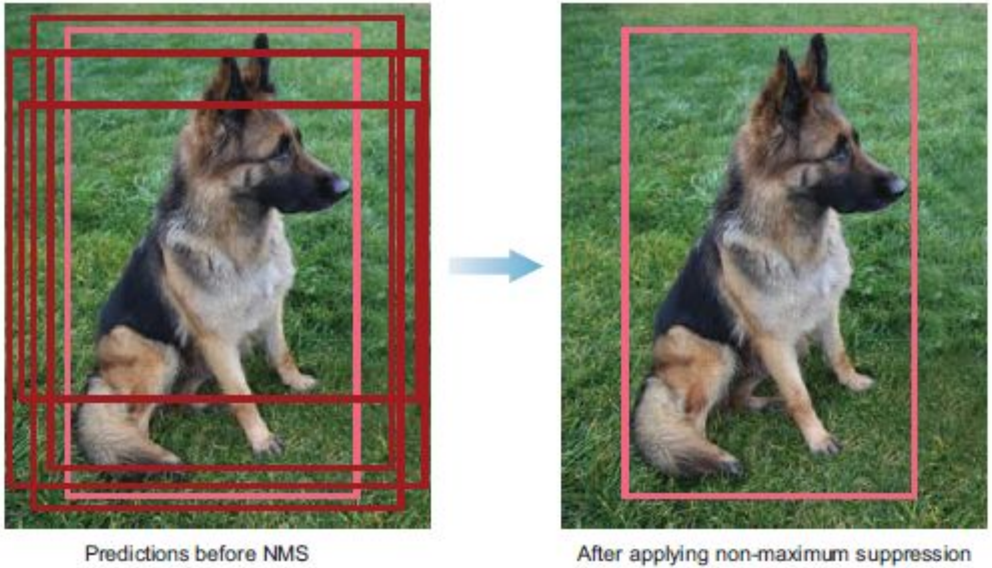
Predictions before NMS — After applying non-maximum suppression

(Elgendy, 2020)

3. Remove Duplicate Object Detections

Non-maximum suppression (NMS):
A 4-step technique for eliminating duplicate detections of objects.

1. Discard BBs with predictions below a **confidence threshold**.
2. Select the BB with the highest probability
3. Calculate the overlap of all remaining boxes with the same class prediction
4. Suppress any box with an IoU smaller than a threshold (NMS threshold, usually 0.5).



(Elgendy, 2020)

4. Evaluation Metrics for Detector Performance

Once an object detector has been developed, it is typically evaluated using two main metrics:

1. **Frames per second (FPS) - detection speed**
2. **Mean Average Precision (mAP) - network precision**

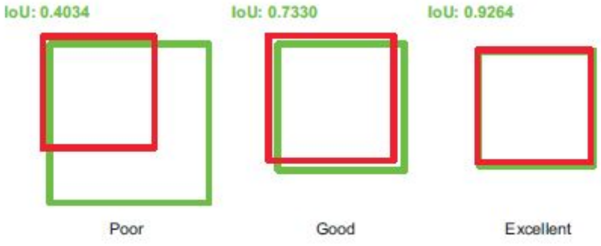
   mAP calculated from **a Bounding Box's object score** and the **precision-recall curve**

(Elgendy, 2020)

4. Evaluation of Detector Performance: IoU

Intersection over Union: IoU

$$\text{IoU} = \frac{B_{\text{ground truth}} \cap B_{\text{predicted}}}{B_{\text{ground truth}} \cup B_{\text{predicted}}}$$



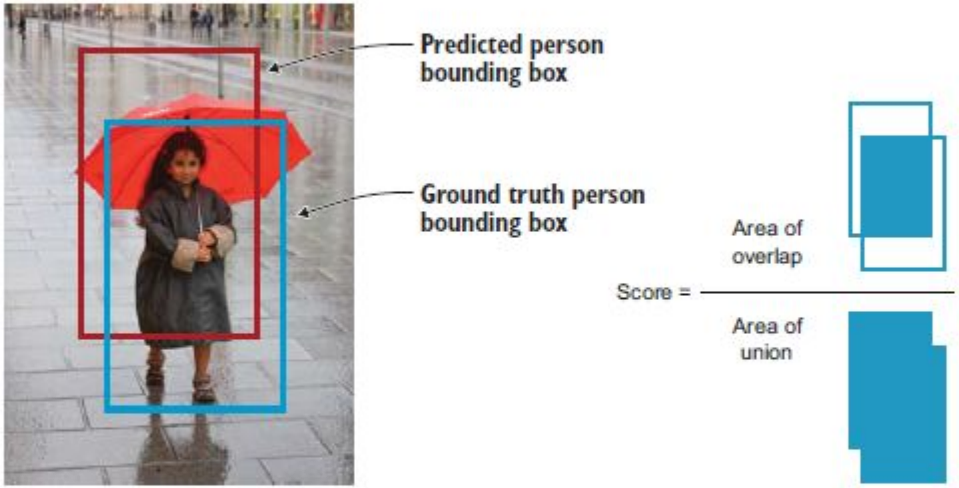


Figure 7.5 The IoU score is the overlap between the ground truth bounding box and the predicted bounding box.

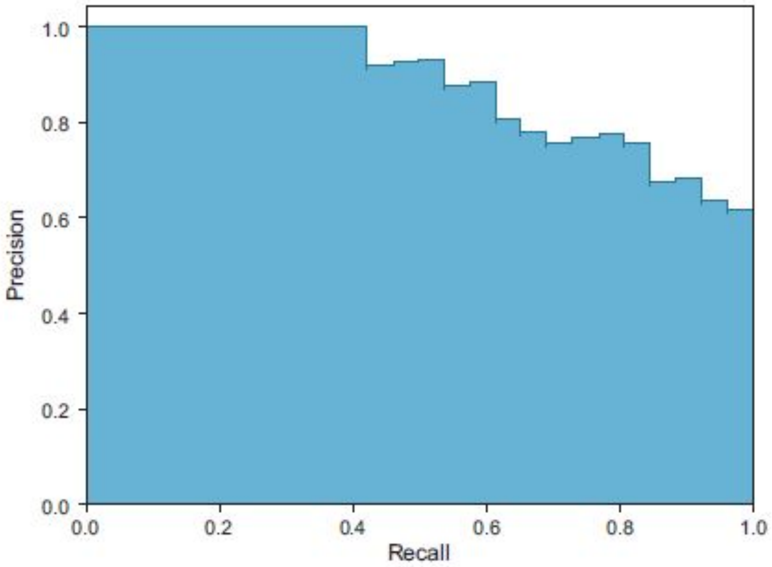
(Elgendy, 2020)

4. Evaluation of Detector Performance: Precision-Recall

Precision-Recall (PR):

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$



(Elgendy, 2020)

State of the Art Object Detection CNNs

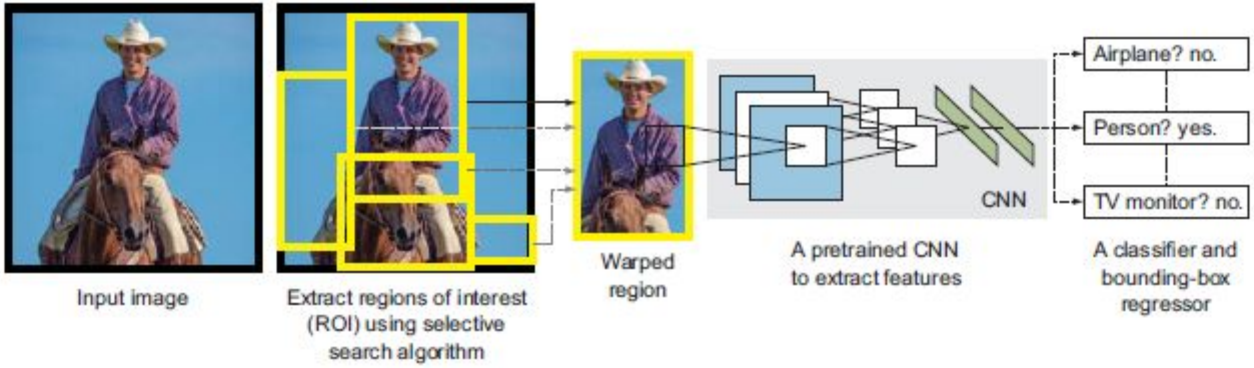
- R-CNNs
- SSD
- YOLO

R-CNNs : Region-based CNNs

R-CNN family of networks :

- **R-CNN**
- Fast-RCNN
- **SOTA:** Faster-RCNN

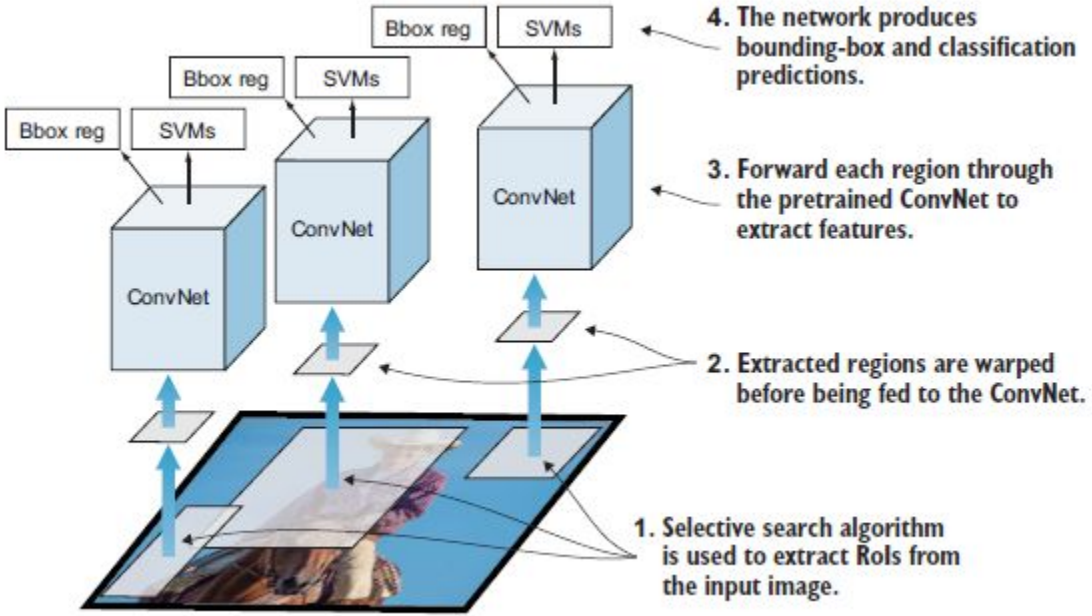
R-CNN architecture ->



(Elgendy, 2020)

R-CNNs : Region-based CNNs

R-CNN :



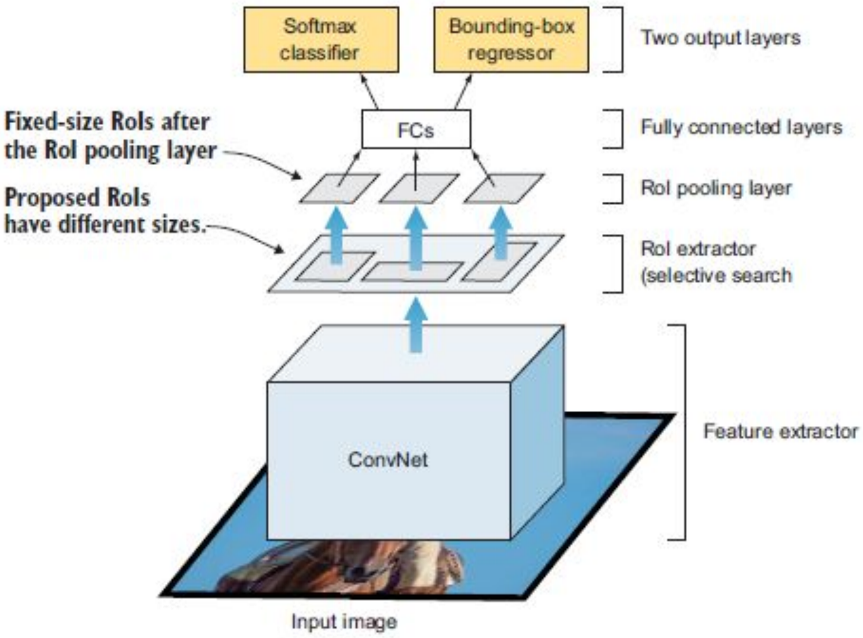
(Elgendy, 2020)

Fast R-CNN

Improves on R-CNN in both detection **speed** and **accuracy**.

Architecture changes:

- CNN feature extractor first applied to entire image, region proposals performed afterward
  - Only run one CNN instead of ~2000 CNNs on overlapping RoIs
- CNN performs **both** the **classification and feature extraction**
  - SVM machine replaced with a softmax layer



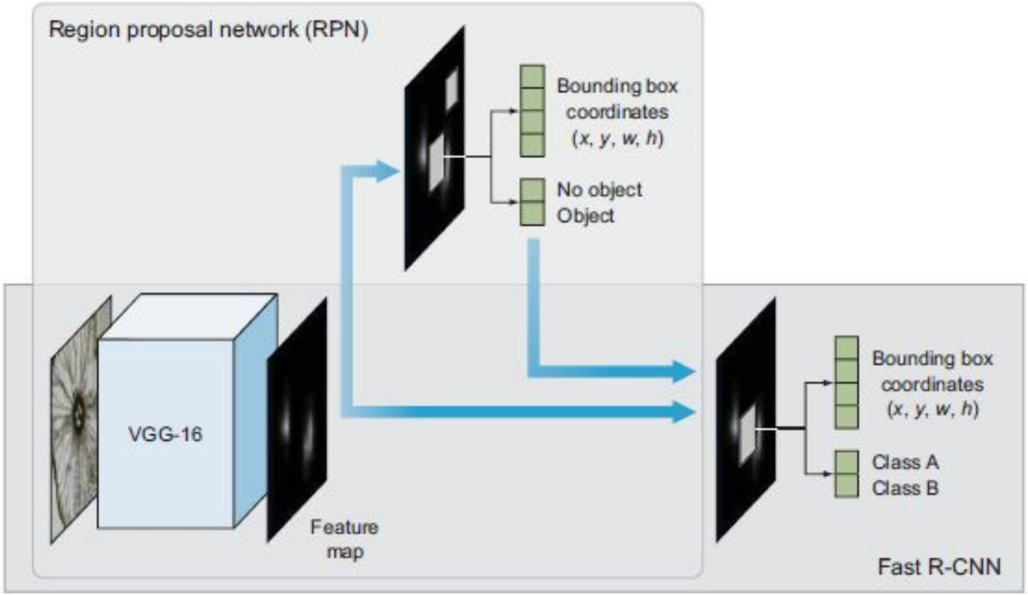
(Elgendy, 2020)

Faster R-CNN: State of the Art

Architecture

- Same overall structure as Fast R-CNN except for Region proposal algorithm

- Selective search replaced with Region proposal network, which Outputs
  - Objectness score
  - Bounding box location



(Elgendy, 2020)

Multi-stage vs Single-Stage Detectors

Multistage detectors:

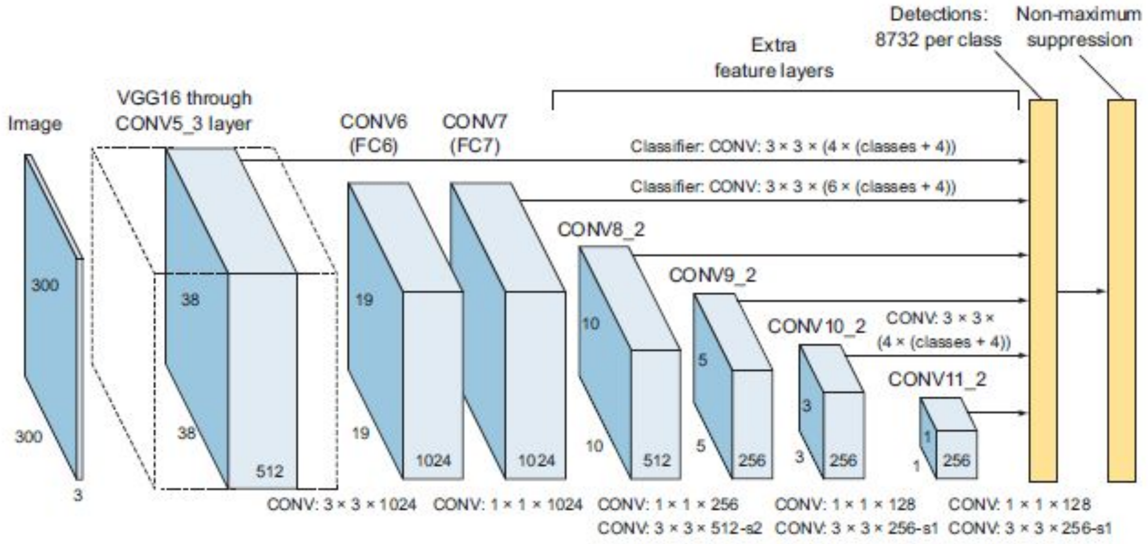
- First Identifies Regions of interest & objective score
- Later outputs final Bounding boxes

Single-stage detectors:

- Identifies Regions of Interest and final bounding boxes together

SSD Single-Shot Detector

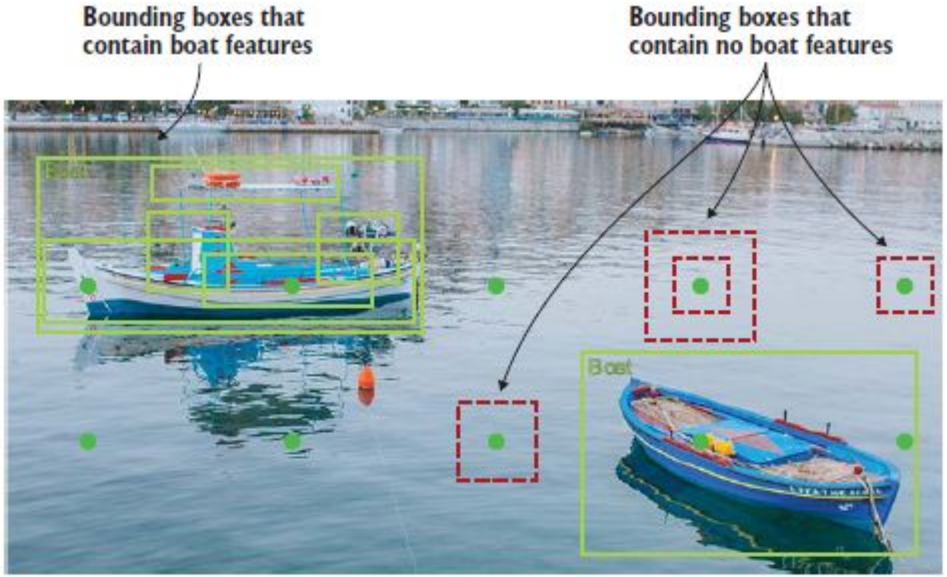
- Pretrained network extracts features; cut off before Classification
- Convolutional layers allow for detections at different scales
- Non-maximum suppression Eliminates overlapping BBs to keep one detection per object
- Outputs a set number of BBs prior to NMS.



(Elgendy, 2020)

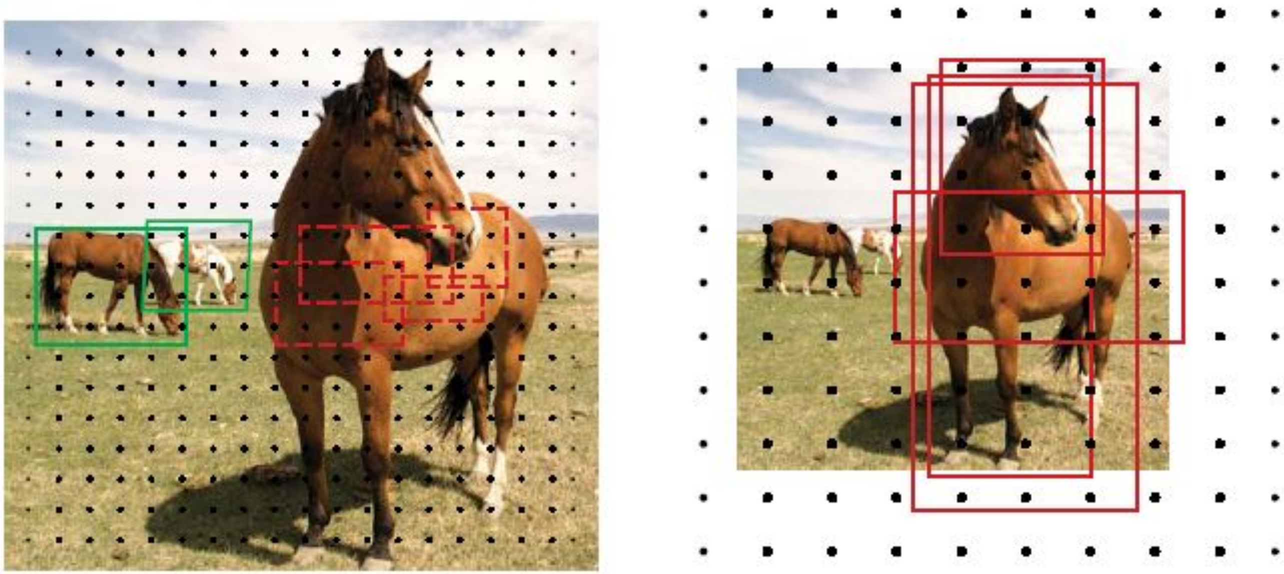
SSD Feature Extraction

- Anchors overlaid over image
- Bounding boxes created with Anchors at their center
- Network considers each BB a separate image
- If feature extractor found boat Features in BB, BB sent on to NMS layer



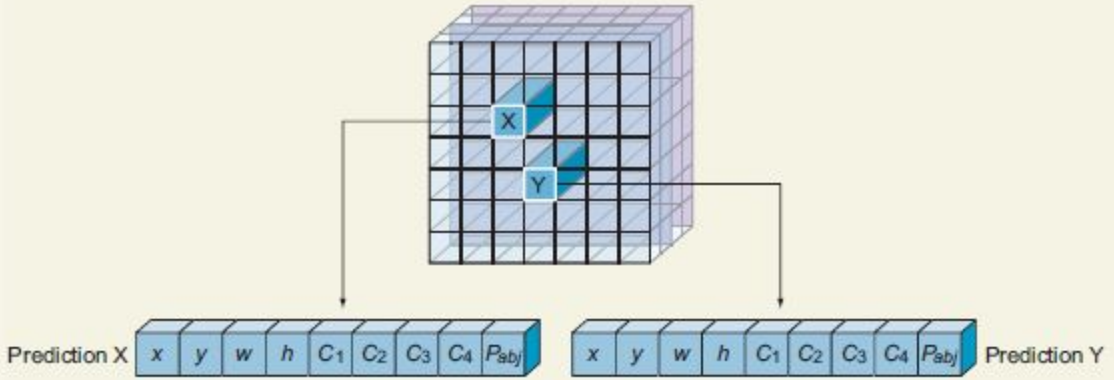
Multi-scale Feature Layers

Convolutional feature layers after the base network



SSD Single-Shot Detector

- Output of SSD

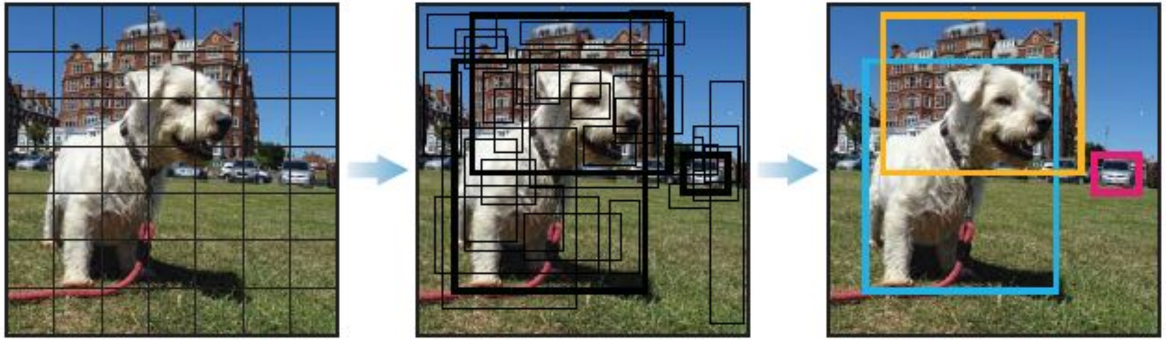


An example visualization of the output prediction when we have four classes in our problem. The convolutional layer predicts the bounding box coordinates, objectness score, and four class probabilities: C_1, C_2, C_3, and C_4.

(Elgendy, 2020)

YOLO: Real-time Object Detection

- No region proposal Network
- Performs predictions Based on a grid of cells
- Each cell directly predicts BBs and classification
- NMS yields final prediction



Splits the image into grids

Predicts bounding boxes and classifications

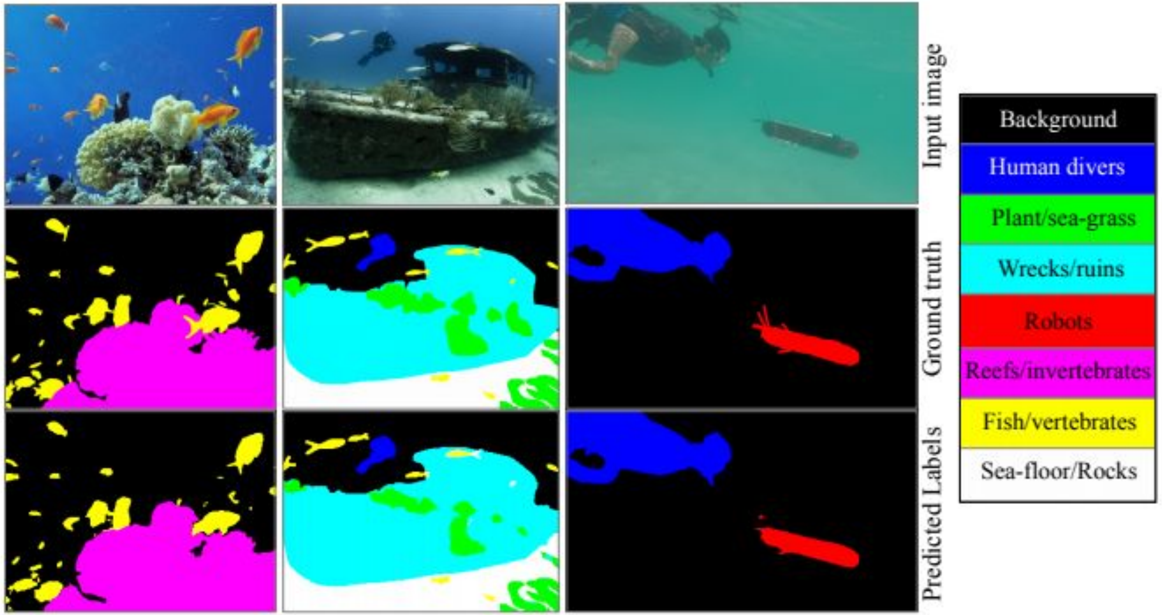
Final predictions after non-maximum suppression

(Elgendy, 2020)

Applications of CNNs in Computer Vision

- Object Detection
- **Segmentation**

Segmentation



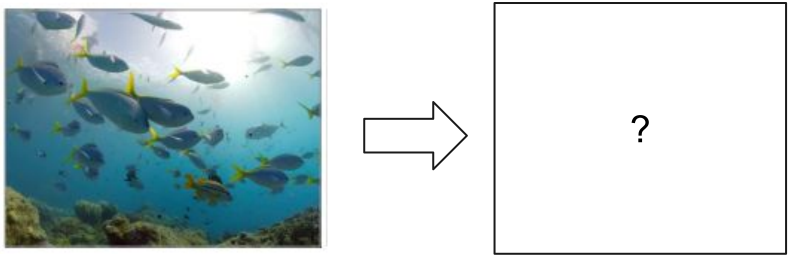
(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)

Paper from UMN IRVLab: http://irvlab.dl.umn.edu/

What is Semantic Segmentation?

FISH, DIVER, BACKGROUND, AQUATIC PLANTS, SEAFLOOR

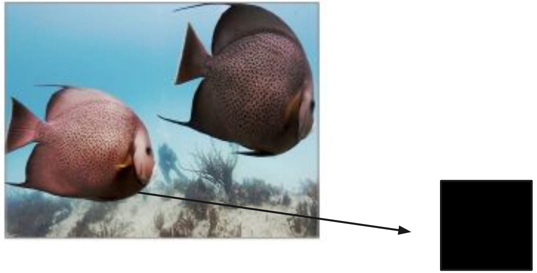
Training data paired:
Each pixel labeled with a semantic category.



During test, classify each pixel of the new image.

(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/

Semantic Segmentation: Sliding Window

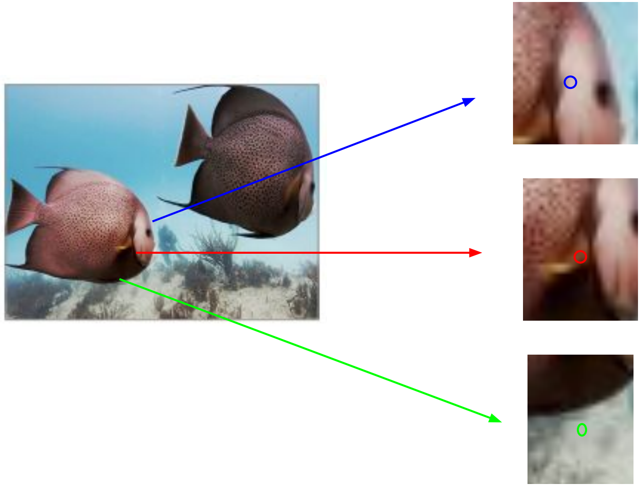


Impossible to classify without context!

How do we include context?

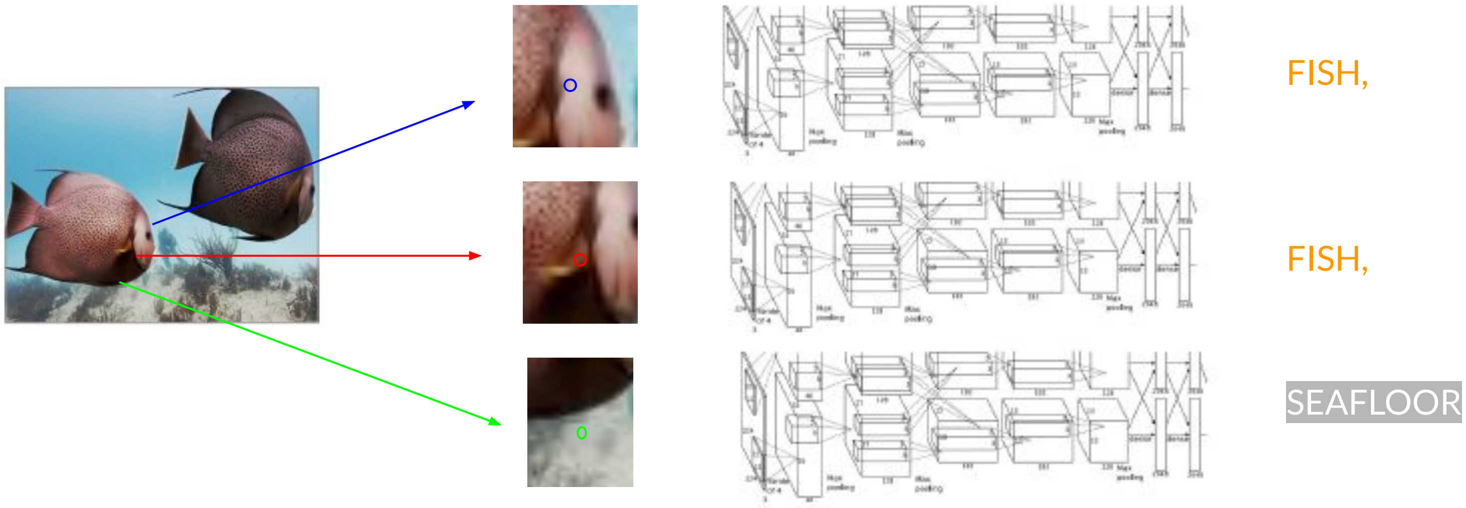
(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/

Semantic Segmentation: Sliding Window



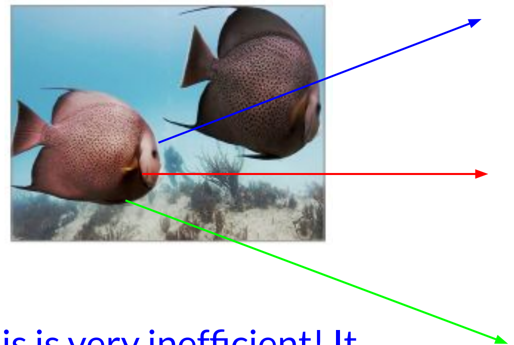
(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

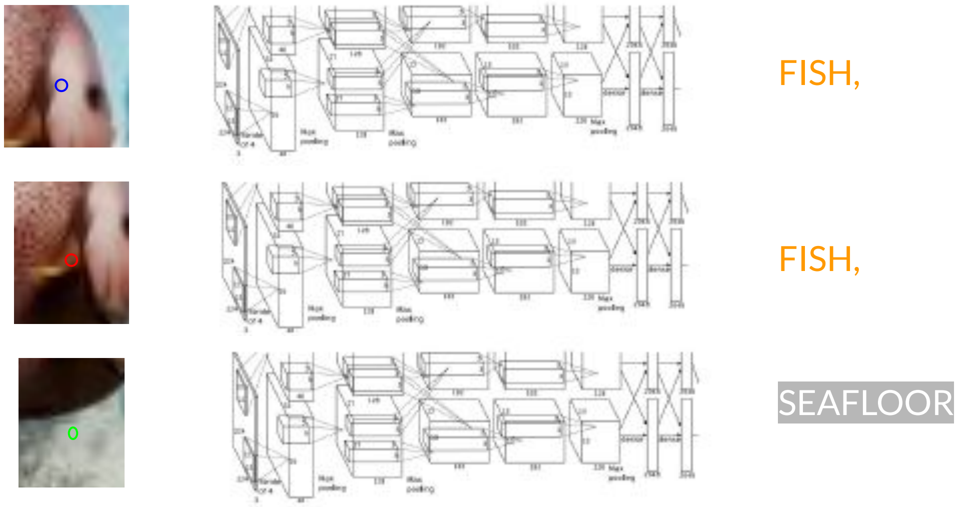
Semantic Segmentation: Sliding Window



(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation: Sliding Window

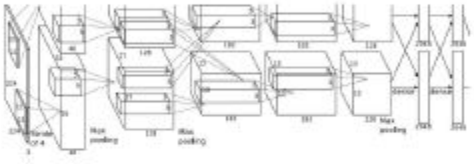




This is very inefficient! It re-identifies shared features for each overlapping patch.

(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation Idea: Convolution



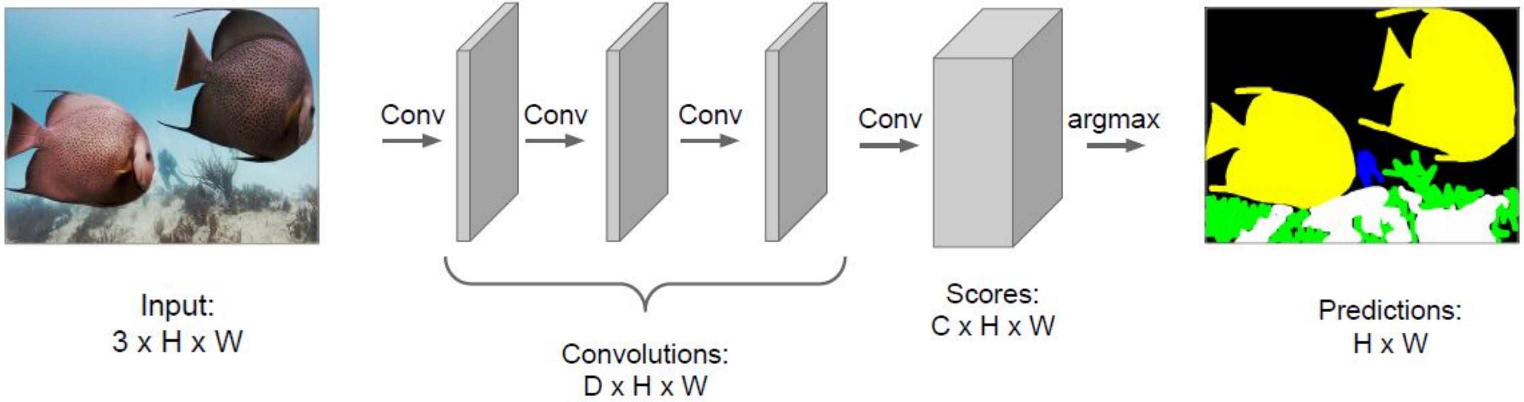
Intuition: encode the entire image with a CNN, then do semantic segmentation at the end.

Challenge: Classification architectures reduce feature sizes as they go deeper into the network; Semantic segmentation requires output size == input.

(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation Idea: Convolution

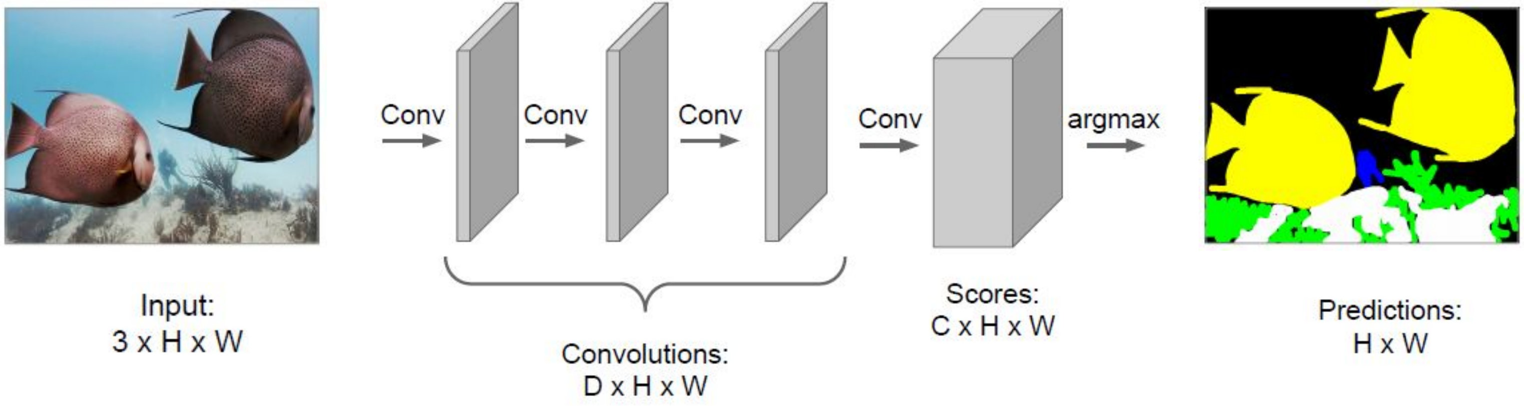
Eliminate downsampling in the network:



(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation Idea: Convolution

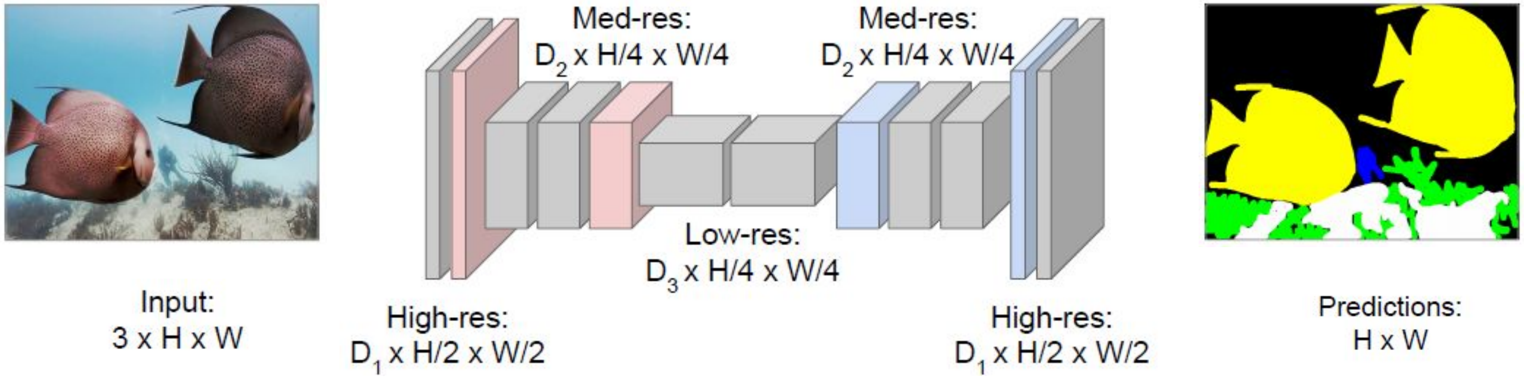
Issue: very computationally expensive!



(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation Idea: Convolution

Solution: add both downsampling and upsampling inside network!

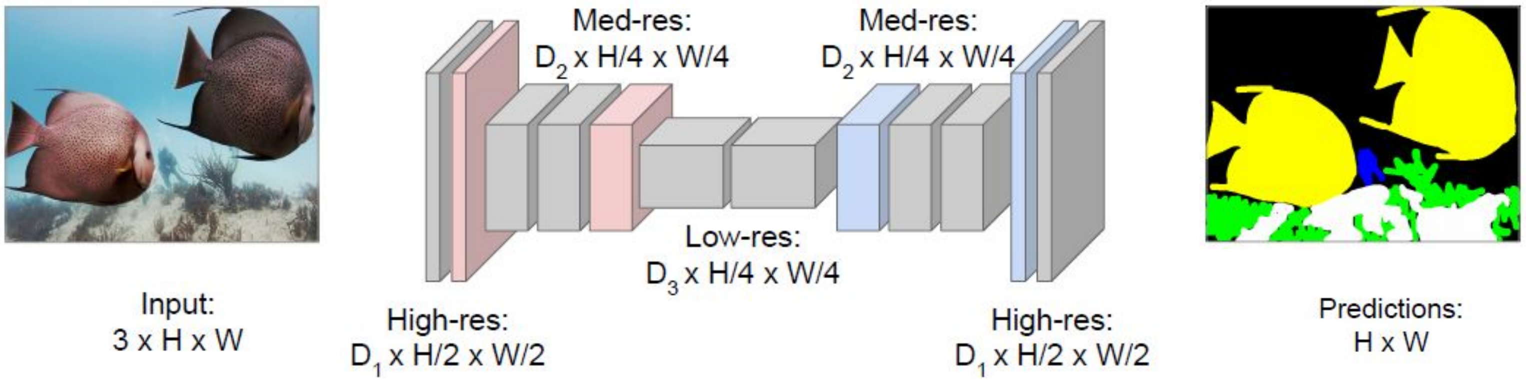


(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation Idea: Convolution

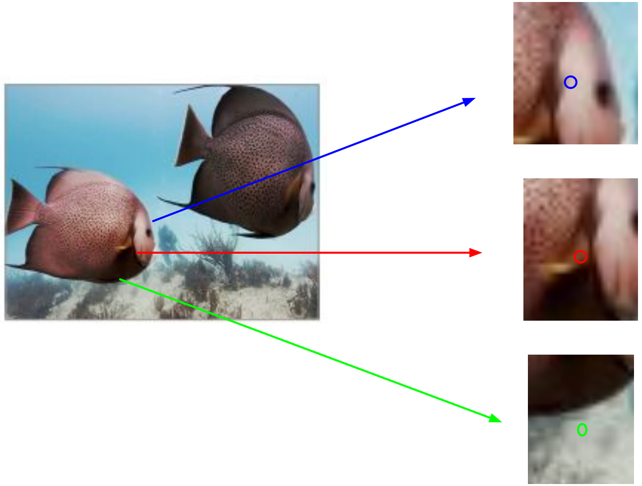
Downsampling:
Pooling, strided convolution

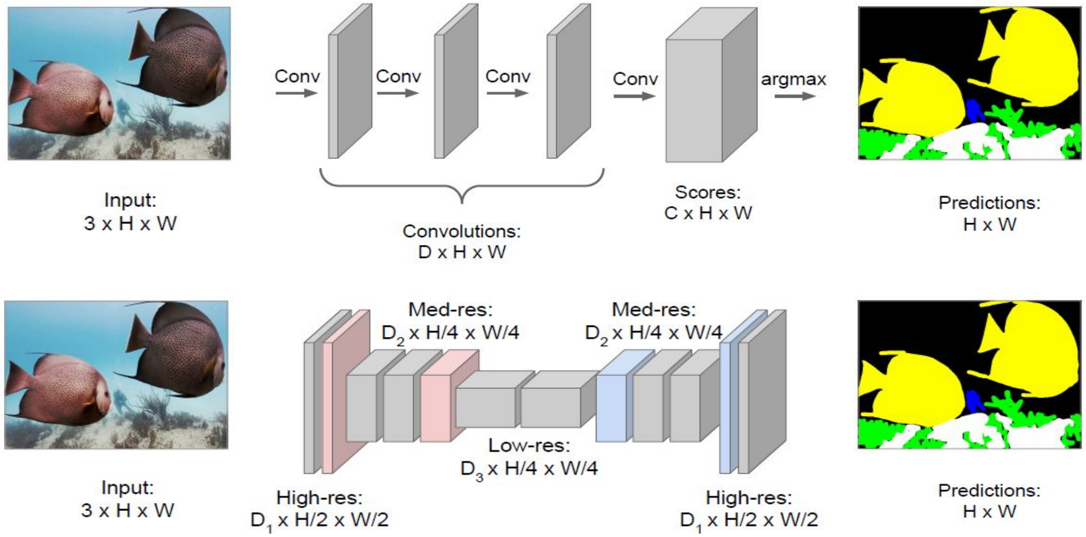
Upsampling: from 1st half of lecture



(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation: Summary



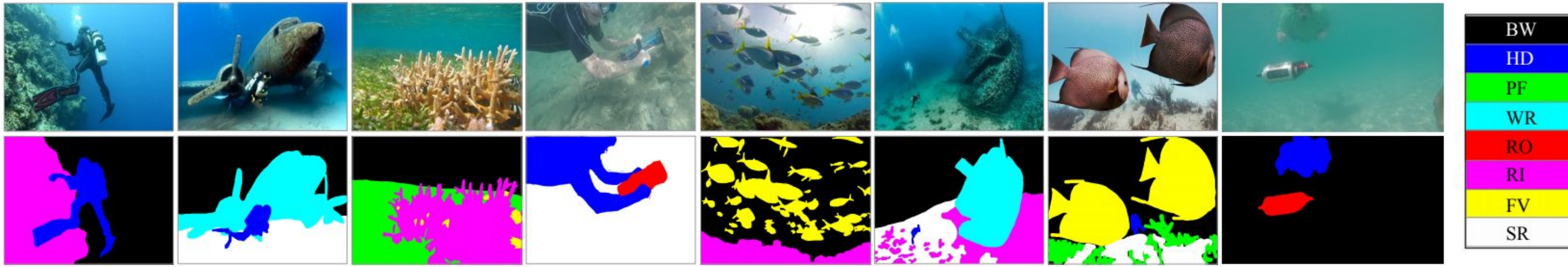


(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/
(Li et al., Detection and Segmentation 2020)

Semantic Segmentation: Summary

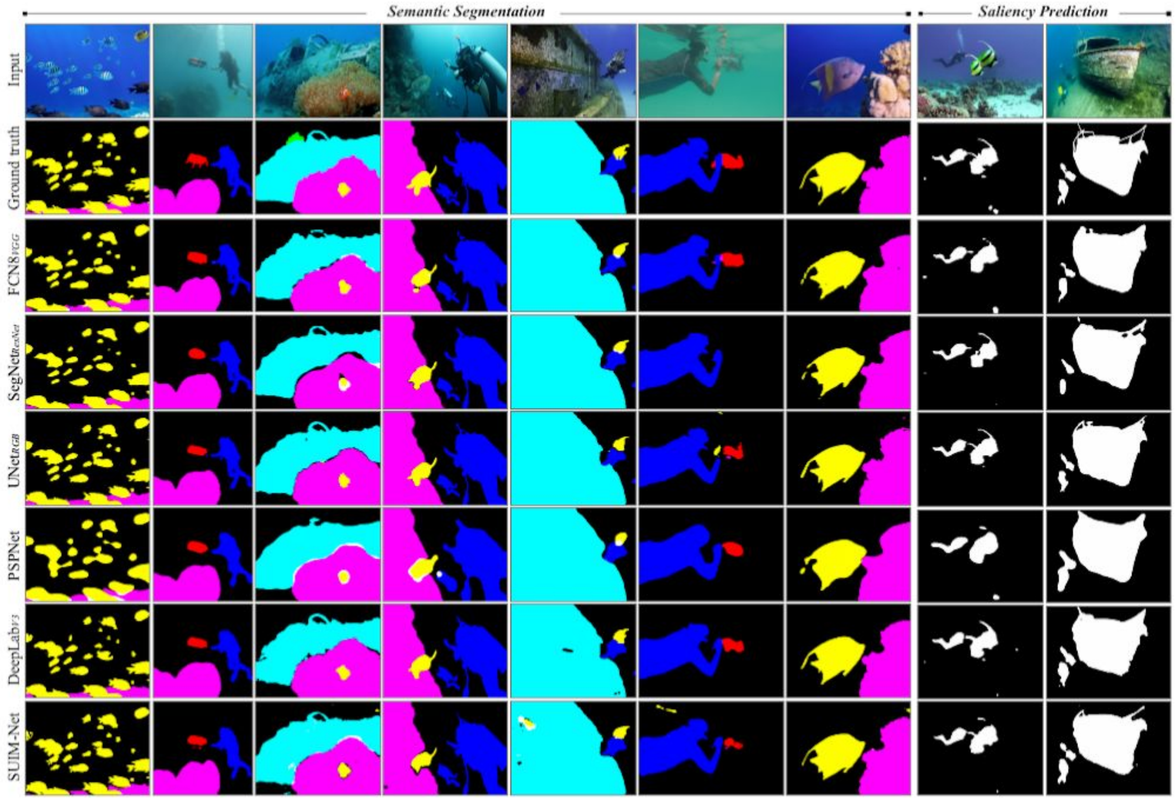
Goal: label each pixel in the image with a category label. Don't differentiate between different instances of the same class of object; only care about the pixel-level.

Object category	RGB color	Code
Background (waterbody)	000	BW
Human divers	001	HD
Aquatic plants and sea-grass	010	PF
Wrecks or ruins	011	WR
Robots (AUVs/ROVs/instruments)	100	RO
Reefs and invertebrates	101	RI
Fish and vertebrates	110	FV
Sea-floor and rocks	111	SR



(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/

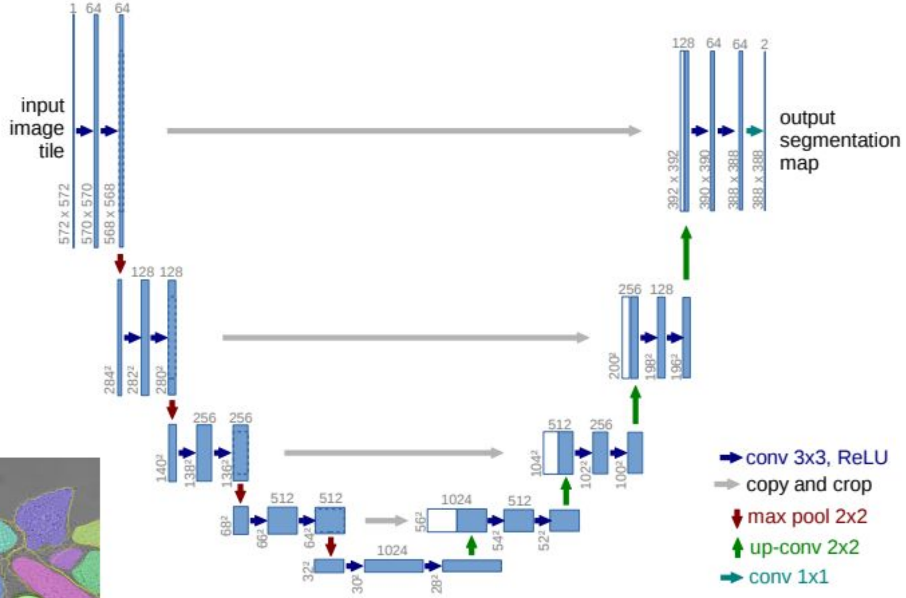
State of the Art Segmentation CNNs

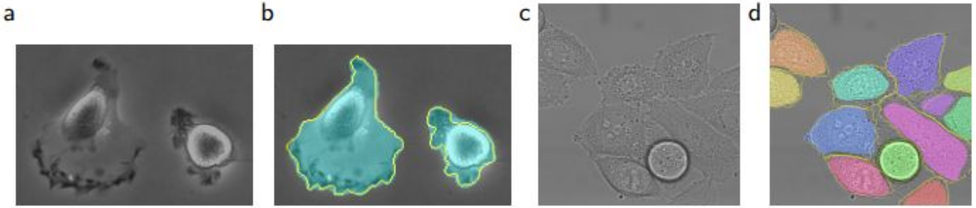


(Islam et al., "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," 2020)
Paper from UMN IRVLab http://irvlab.dl.umn.edu/

UNET: Semantic Segmentation

- Applied to medical imaging to identify Tumors
- Main contribution: addition of “Shortcut connections:
- Compensates for information loss, Since not all features can be re-created By the decoder





(Ronneberger et al., 2015)

Mask R-CNN: Instance Segmentation

- Extension of Faster R-CNN
- Adds a masking network after the Output of Faster R-CNN
- Masking network outputs a Segmentation mask for each Object instance

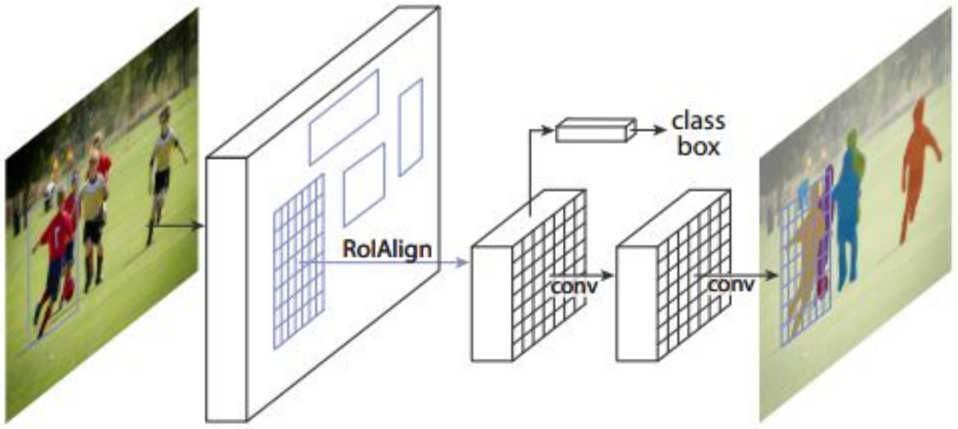


Figure 1. The **Mask R-CNN** framework for instance segmentation.

(He et al., 2018)

Mask R-CNN: Instance Segmentation

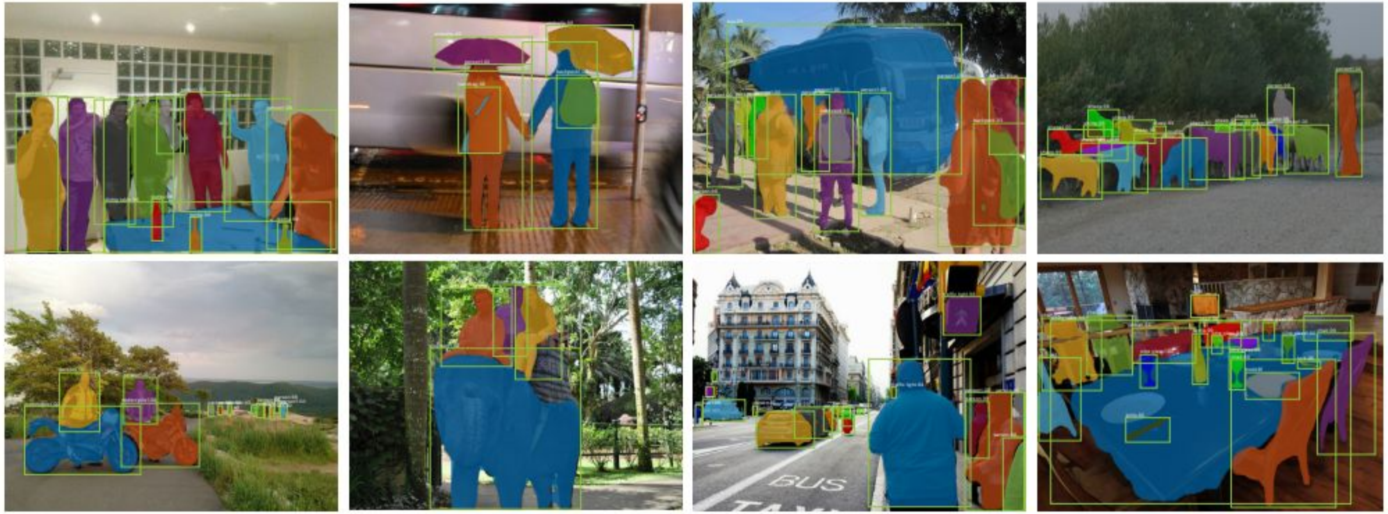


Figure 2. **Mask R-CNN** results on the COCO test set. These results are based on ResNet-101 [19], achieving a *mask* AP of 35.7 and running at 5 fps. Masks are shown in color, and bounding box, category, and confidences are also shown.

(He et al., 2018)

Popular Datasets for Classification, Detection, and Segmentation

- COCO (172 classes, common benchmark dataset)
  - http://cocodataset.org/#home
- Cityscapes (roads, lanes vehicles, objects on roads)
  - https://www.cityscapes-dataset.com/
- Pascal Context (real-world; over 400 classes)
  - https://cs.stanford.edu/~roozbeh/pascal-context/
- Lits (medical imaging, CT scans)
  - https://competitions.codalab.org/competitions/17094
- Inria Aerial Image Labeling

Acknowledgements

F.-F. Li, R. Krishna, and D. Xu, “Detection and Segmentation,” in *CS231n: Convolutional Neural Networks for Visual Recognition*, 2020.

F.-F. Li, R. Krishna, and D. Xu, “Visualizing and Understanding,” in *CS231n: Convolutional Neural Networks for Visual Recognition*, 2020.

M. J. Islam, M. Fulton and J. Sattar, "Toward a Generic Diver-Following Algorithm: Balancing Robustness and Efficiency in Deep Visual Detection," in IEEE Robotics and Automation Letters, vol. 4, no. 1, pp. 113-120, Jan. 2019, doi: 10.1109/LRA.2018.2882856.

M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, “Semantic Segmentation of Underwater Imagery: Dataset and Benchmark,” in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE/RSJ, 2020.

M. Elgendy, “Object Detection with R-CNN, SSD, and YOLO,” in *Deep Learning for Vision Systems*, Shelter Island, NY: O'REILLY MEDIA, 2020, pp. 283–337.

K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” 2018.

O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,”2015.