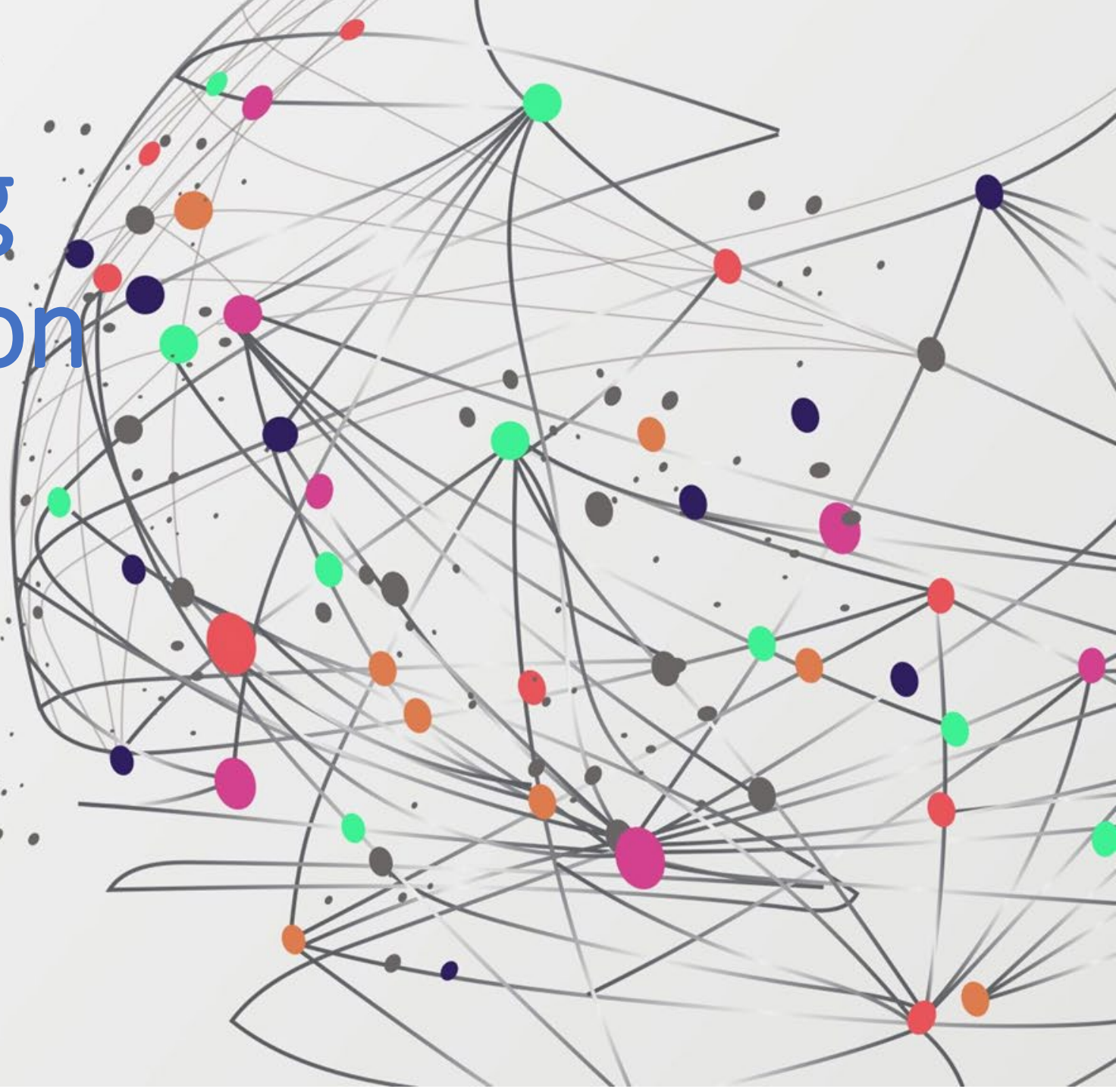
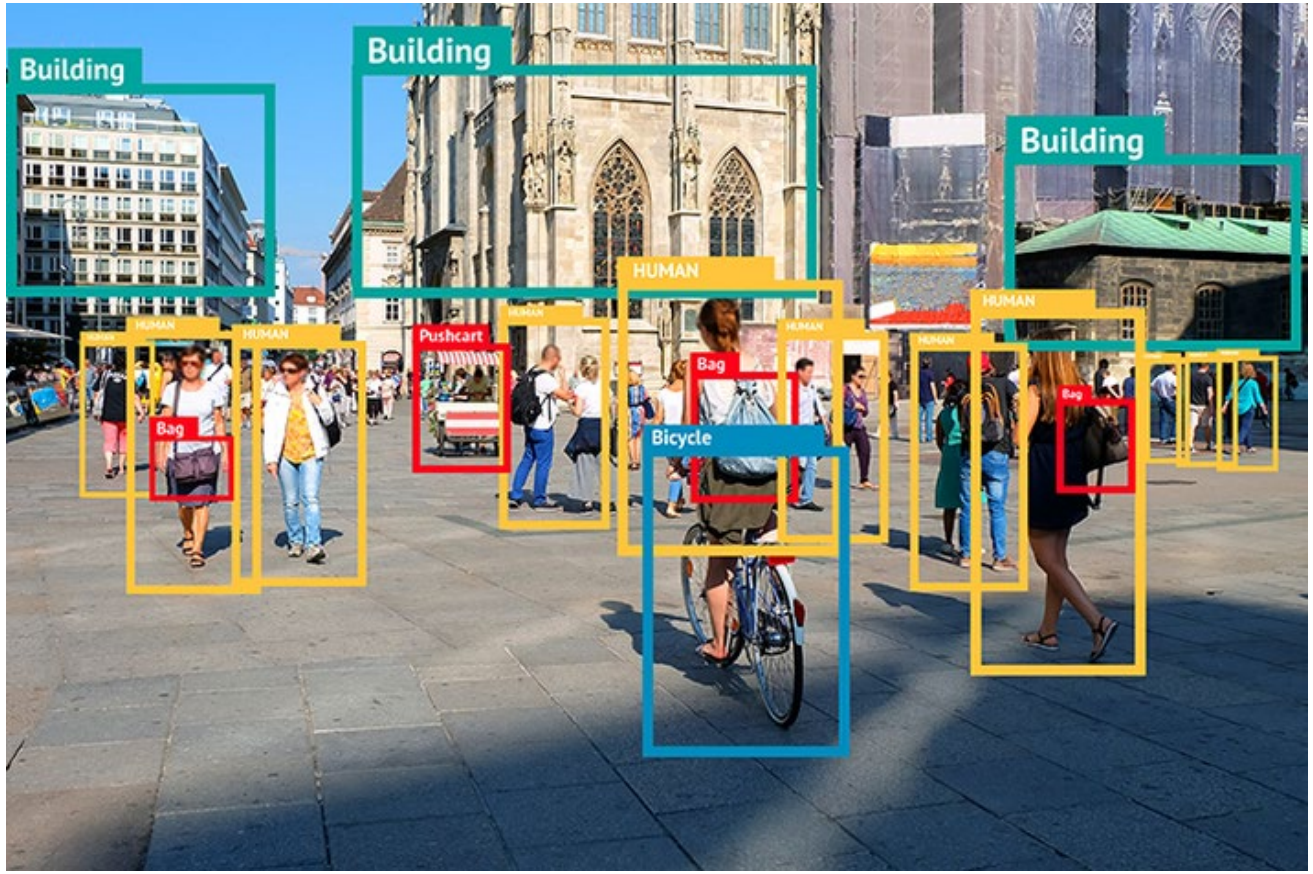
Deep Learning Object Detection & Semantic Segmentation



Silverio García Cortés
Universidad de Oviedo
Naples. May 2023, Univ. Federico II

Deep Learning: Object Detection and Object recognition

- Image classification involves assigning a class label to an image, whereas object localization involves drawing a bounding box around one or more objects in an image and label them. **Object detection** is more challenging and combines these two tasks: draws a bounding box around each object of interest in the image and assigns them a label.



- **Image classification is a self explained expression, but the differences between object localization and object detection can be confusing, especially when all three tasks may be just as equally referred to as object recognition.**

- **Object recognition refers to a collection of related tasks for identifying objects in digital photographs.**

- **Region-Based Convolutional Neural Networks, or R-CNNs, are a family of techniques for addressing object localization and recognition tasks, designed for model performance.**

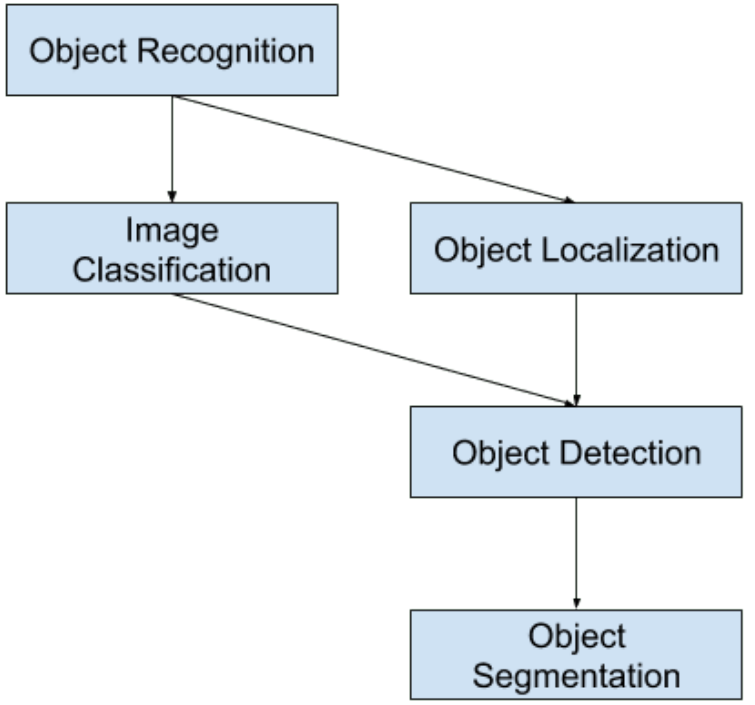
- **You Only Look Once, or YOLO, is a second family of techniques for object recognition designed for speed and real-time use.**

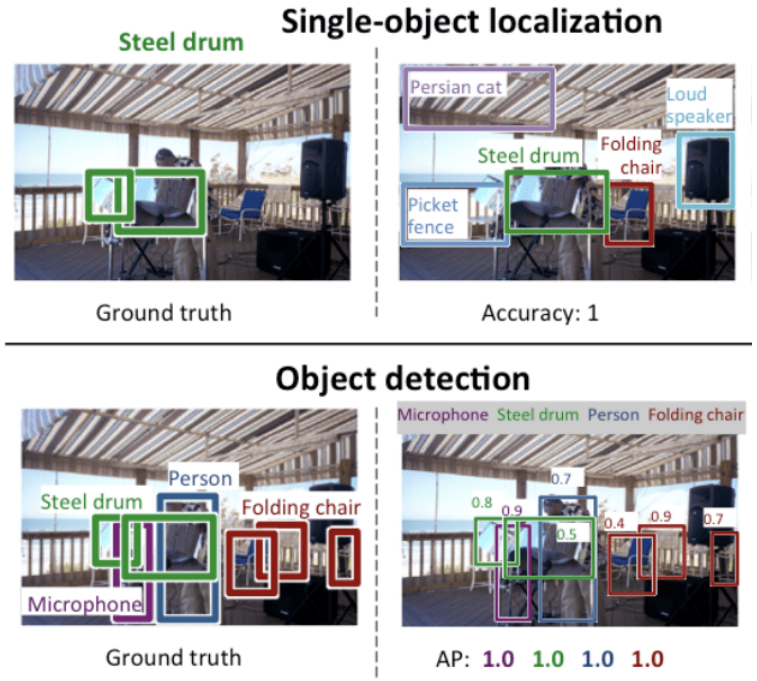
Object Detection & Object segmentation

- **Image Classification**: Predict the type or class of an object in an image.
  - *Input*: An image with a single object, such as a photograph.
  - *Output*: A class label (e.g. one or more integers that are mapped to class labels).
- **Object Localization**: Locate the presence of objects in an image and indicate their location with a bounding box.
  - *Input*: An image with one or more objects, such as a photograph.
  - *Output*: One or more bounding boxes (e.g. defined by a point, width, and height).
- **Object Detection**: Locate the presence of objects with a bounding box and types or classes of the located objects in an image.
  - *Input*: An image with one or more objects, such as a photograph.
  - *Output*: One or more bounding boxes (e.g. defined by a point, width, and height), and a class label for each bounding box.

Object segmentation :

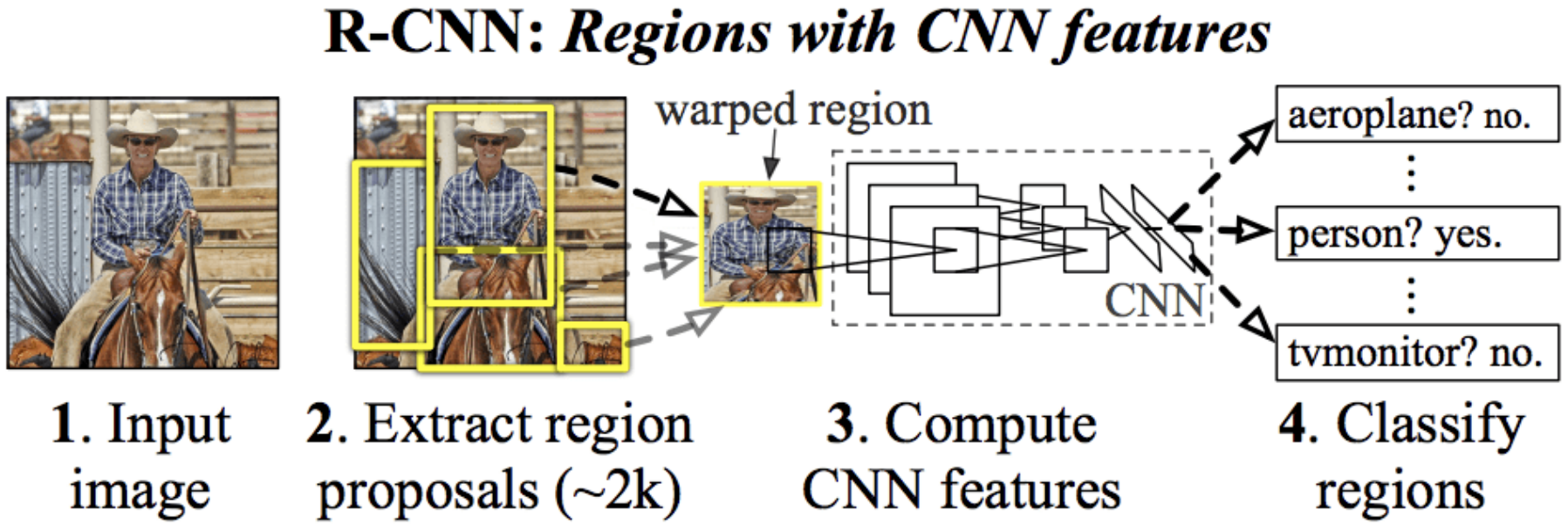
- One further extension to this breakdown of computer vision tasks is object segmentation, also called “object instance segmentation” or “semantic segmentation,” where instances of recognized objects are indicated by highlighting the specific pixels of the object instead of a coarse bounding box.



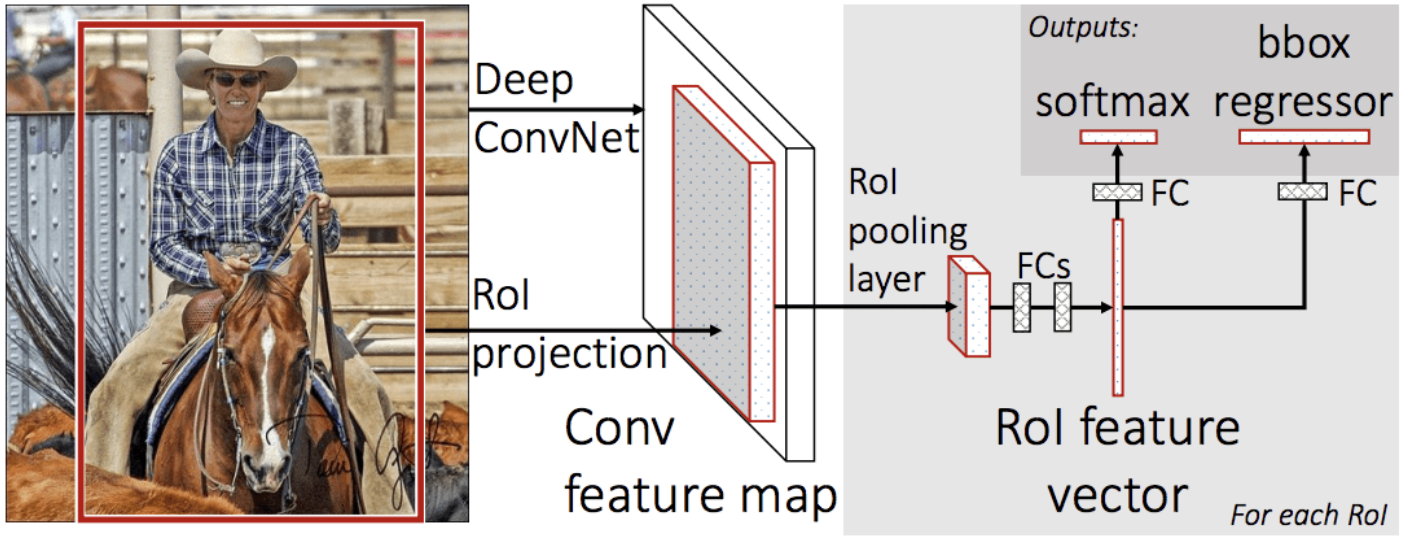


R-CNN. Region Based Convolutional Neural Network

- 2014. Ross Girshick, et al. from UC Berkeley titled "Rich feature hierarchies for accurate object detection and semantic segmentation.
- Their proposed R-CNN model is comprised of three modules:
  - **Module 1: Region Proposal.** Generate and extract category independent region proposals, e.g. candidate bounding boxes.
  - **Module 2: Feature Extractor**. Extract feature from each candidate region, e.g. using a deep convolutional neural network.
  - **Module 3: Classifier**. Classify features as one of the known class, e.g. linear SVM classifier model.
- 2. A computer vision technique is used to propose candidate regions or bounding boxes of potential objects in the image called "*selective search*," although the flexibility of the design allows other region proposal algorithms to be used.
- 3. The feature extractor used by the model was the AlexNet deep CNN that won the ILSVRC-2012 image classification competition
- 4. The output of the CNN was a 4,096 element vector that describes the contents of the image that is fed to a linear SVM (Support Vector Machines) for classification, specifically one SVM is trained for each known class
- A downside of the approach is that it is slow, requiring a CNN-based feature extraction pass on each of the candidate regions generated by the region proposal algorithm. This is a problem as the paper describes the model operating upon approximately 2,000 proposed regions per image at test-time.



Fast R-CNN



Ross Girshick, then at Microsoft Research, proposed an extension to address the speed issues of R-CNN in a 2015 paper titled "Fast R-CNN."

- The paper opens with a review of the limitations of R-CNN, which can be summarized as follows:
  - **Training is a multi-stage pipeline**. Involves the preparation and operation of three separate models.
  - **Training is expensive in space and time**. Training a deep CNN on so many region proposals per image is very slow.
  - **Object detection is slow**. Make predictions using a deep CNN on so many region proposals is very slow.

- Fast R-CNN is proposed as a single model instead of a pipeline to learn and output regions and classifications directly.

- The architecture of the model takes the photograph a set of region proposals as input that are passed through a deep convolutional neural network.
- A pre-trained CNN, such as a VGG-16, is used for feature extraction.
- The end of the deep CNN is a custom layer called a Region of Interest Pooling Layer, or RoI Pooling, that extracts features specific for a given input candidate region.
- The output of the CNN is then interpreted by a fully connected layer then the model bifurcates into two outputs, one for the class prediction via a softmax layer, and another with a linear output for the bounding box.
- This process is then repeated multiple times for each region of interest in a given image.
- The model is significantly faster to train and to make predictions, yet still requires a set of candidate regions to be proposed along with each input image.

Faster R-CNN

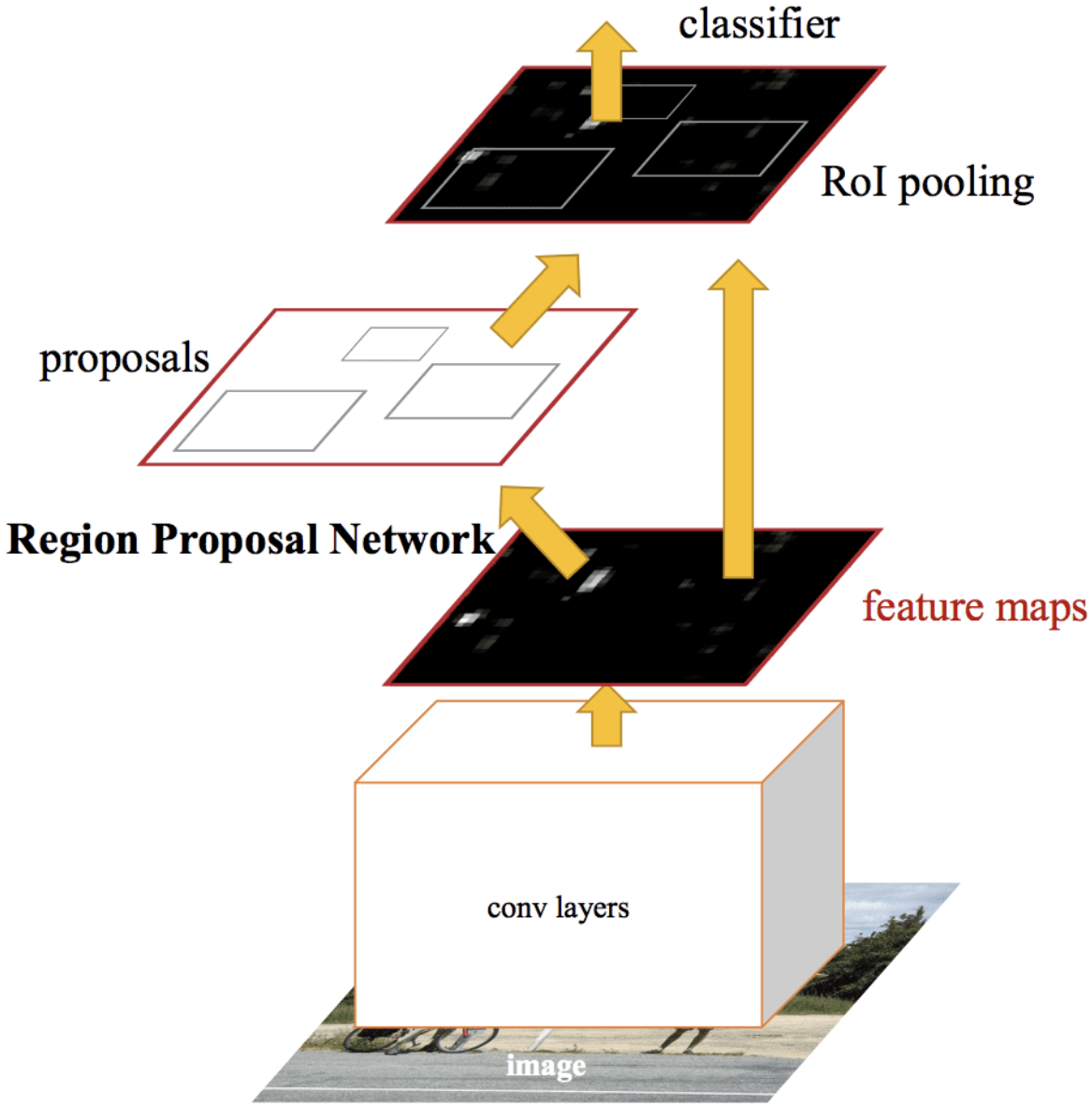
- 2016 Shaoqing Ren, et al.“Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.”

- The architecture was designed to both propose and refine region proposals as part of the training process, referred to as a Region Proposal Network, or RPN.
- These regions are then used in concert with a Fast R-CNN model in a single model design
- These improvements both reduce the number of region proposals and accelerate the test-time operation of the model to near real-time with then state-of-the-art performance.

The architecture is comprised of two modules:

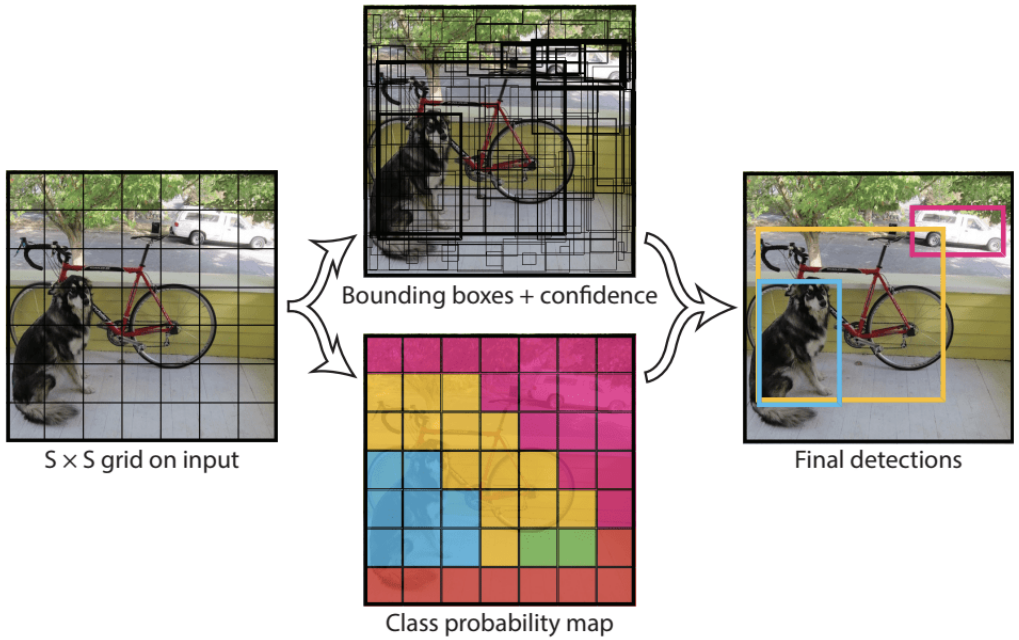
- **Module 1: Region Proposal Network**. Convolutional neural network for proposing regions and the type of object to consider in the region.
- **Module 2: Fast R-CNN**. Convolutional neural network for extracting features from the proposed regions and outputting the bounding box and class labels.

- Both modules operate on the same output of a deep CNN. The region proposal network acts as an attention mechanism for the Fast R-CNN network, informing the second network of where to look or pay attention.
- The RPN works by taking the output of a pre-trained deep CNN, such as VGG-16, and passing a small network over the feature map and outputting multiple region proposals and a class prediction for each.
- Region proposals are bounding boxes, based on so-called anchor boxes or pre-defined shapes designed to accelerate and improve the proposal of regions.
- The class prediction is binary, indicating the presence of an object, or not, so-called “*objectness*” of the proposed region.



YOLO Model Family

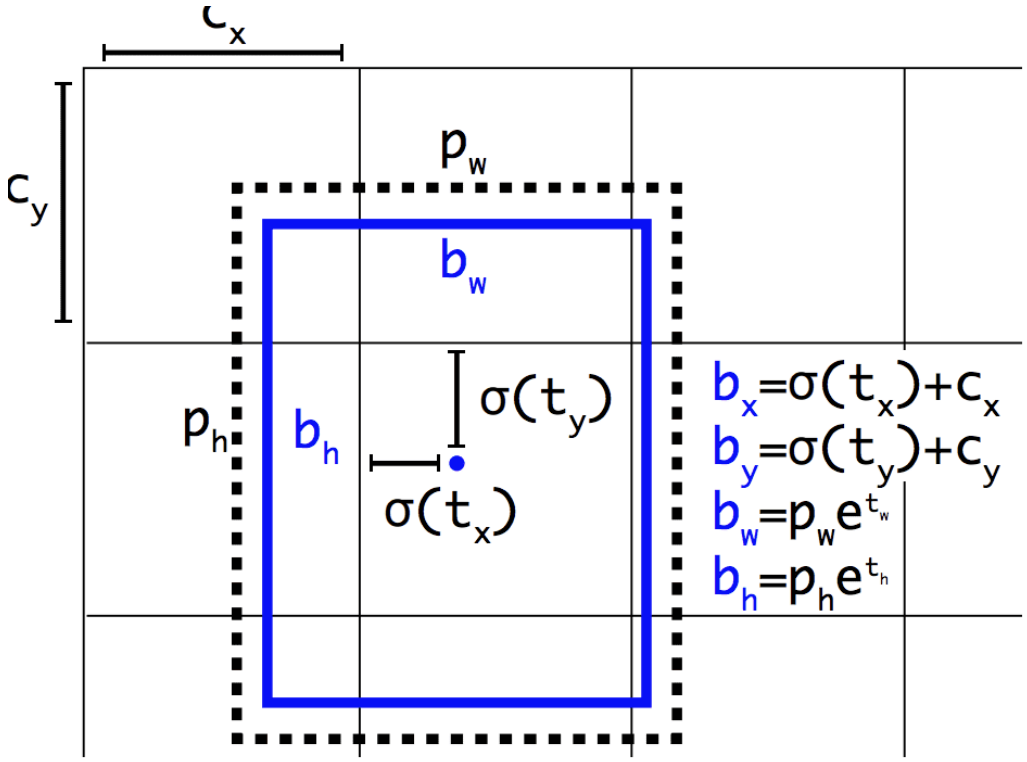
- "You Only Look Once: Unified, Real-Time Object Detection" developed by Joseph Redmon, et al..2016
- The R-CNN models may be generally more accurate, yet the YOLO family of models are fast, much faster than R-CNN, achieving object detection in real-time..
- The approach involves a single neural network trained end to end that takes a photograph as input and predicts bounding boxes and class labels for each bounding box directly.
- The technique offers lower predictive accuracy (e.g. more localization errors), although operates at 45 frames per second and up to 155 frames per second for a speed-optimized version of the model.
- The model works by first splitting the input image into a grid of cells, where each cell is responsible for predicting a bounding box if the center of a bounding box falls within the cell.
- Each grid cell predicts a bounding box involving the x, y coordinate and the width and height and the confidence. A class prediction is also based on each cell.
- For example, an image may be divided into a 7×7 grid and each cell in the grid may predict 2 bounding boxes, resulting in 94 proposed bounding box predictions. The class probabilities map and the bounding boxes with confidences are then combined into a final set of bounding boxes and class labels. The image taken from the paper below summarizes the two outputs of the model.



YOLOv2 (YOLO9000) and YOLOv3

- 2016. Joseph Redmon and Ali Farhadi "YOLO9000: Better, Faster, Stronger."

- An instance of the model is described that was trained on two object recognition datasets in parallel, capable of predicting 9,000 object classes
- Some changes include the use of batch normalization and high-resolution input images.
- YOLOv2 model makes use of anchor boxes, pre-defined bounding boxes with useful shapes and sizes that are tailored during training.
- Rather than predicting position and size directly, offsets are predicted for moving and reshaping the pre-defined anchor boxes relative to a grid cell and dampened by a logistic function.
- Further improvements to the model were proposed by Joseph Redmon and Ali Farhadi in their 2018 paper titled *"YOLOv3: An Incremental Improvement."* The improvements were reasonably minor, including a deeper feature detector network and minor representational changes.



Some examples

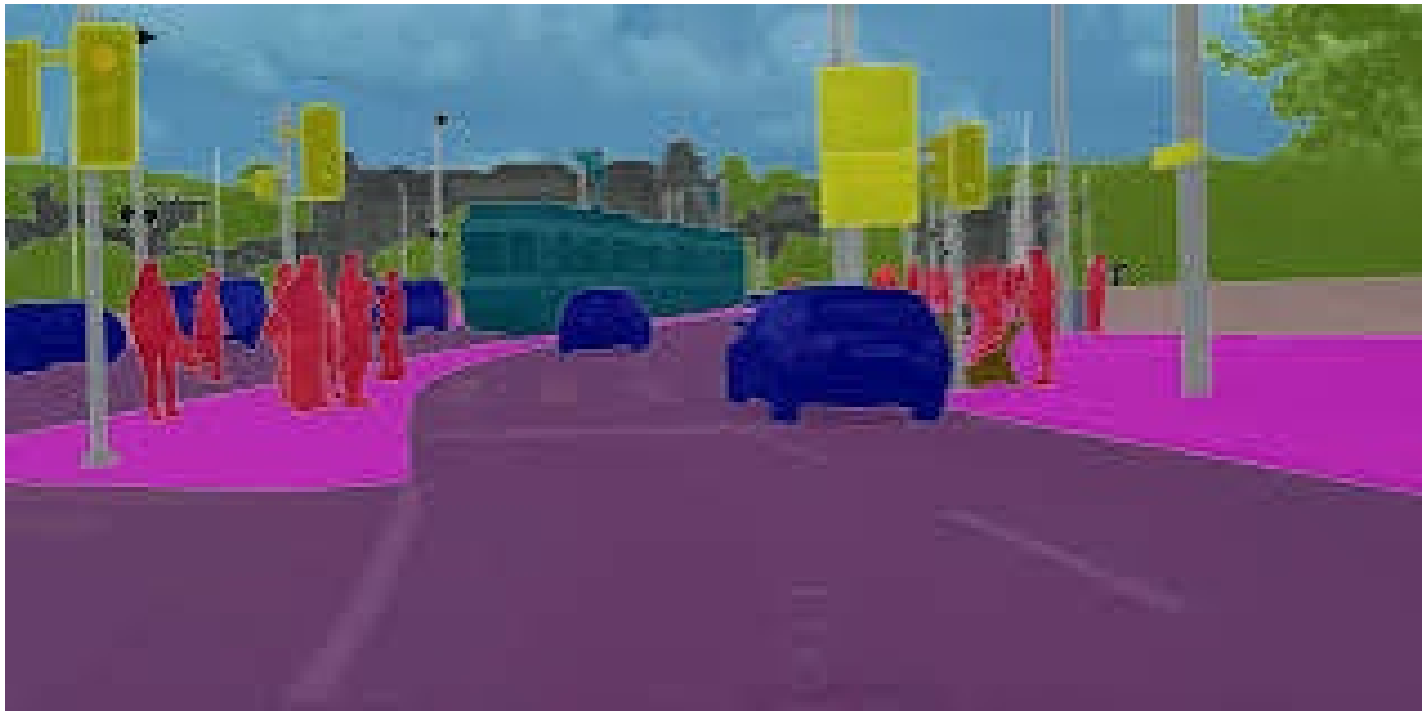
https://youtu.be/lVdh571SwOQ

Semantic Segmentation

Semantic segmentation

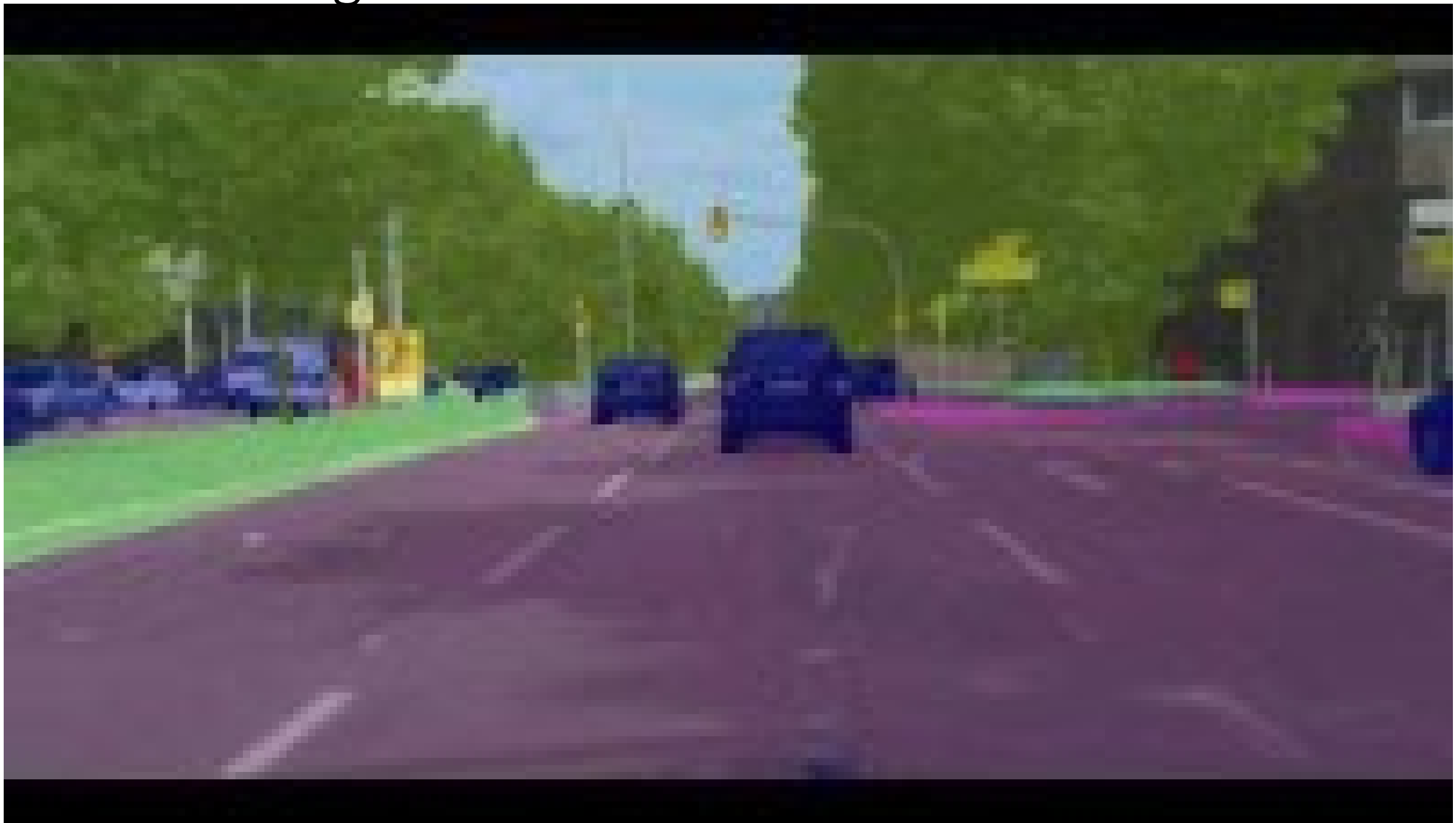
- Semantic segmentation takes the basic task of image classification one step further. Image classification involves assigning a label to an entire image (for example, identifying that it is an image of a dog, cat, or horse). However, naive image classification is limited in real-world computer vision applications, because most images contain more than one object.
- This creates the need to divide an image into regions, and classify each region separately. The process of dividing the image into regions is called segmentation.

Semantic Segmentation is a computer vision task in which the goal is to categorize each pixel in an image into a class or object.
The goal is to produce a dense pixel-wise segmentation map of an image, where each pixel is assigned to a specific class or object.



Semantic segmentation video

https://youtu.be/0yCYro9H1kM



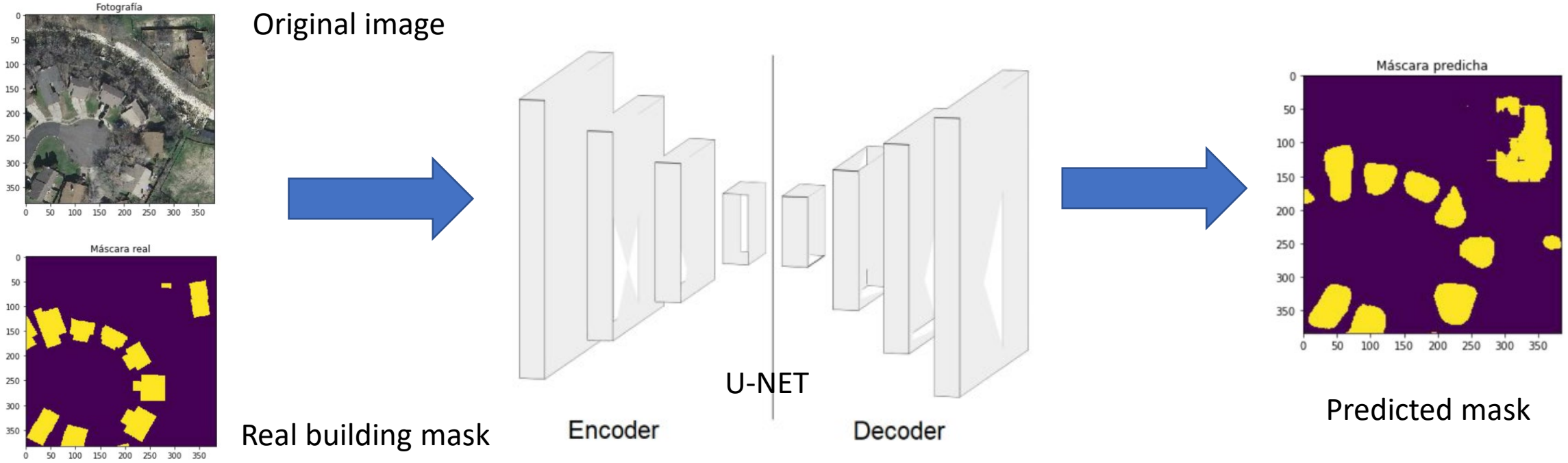
Semantic Segmentation vs. Instance Segmentation

- There are two main types of image segmentation, both of which form the basis for object recognition in computer vision projects:
  - **Semantic segmentation**—objects displayed in the image are grouped according to predefined classes or categories. For example, a city scene can be divided into pedestrians, vehicles, roads, buildings, etc.
  - **Instance segmentation**—identifies specific entities within a class. For example, if semantic segmentation identifies all the pedestrians in the image, instance segmentation can identify individual pedestrians.

Object Detection | Semantic Segmentation | Instance Segmentation

Deep Learning Semantic Segmentation Methods

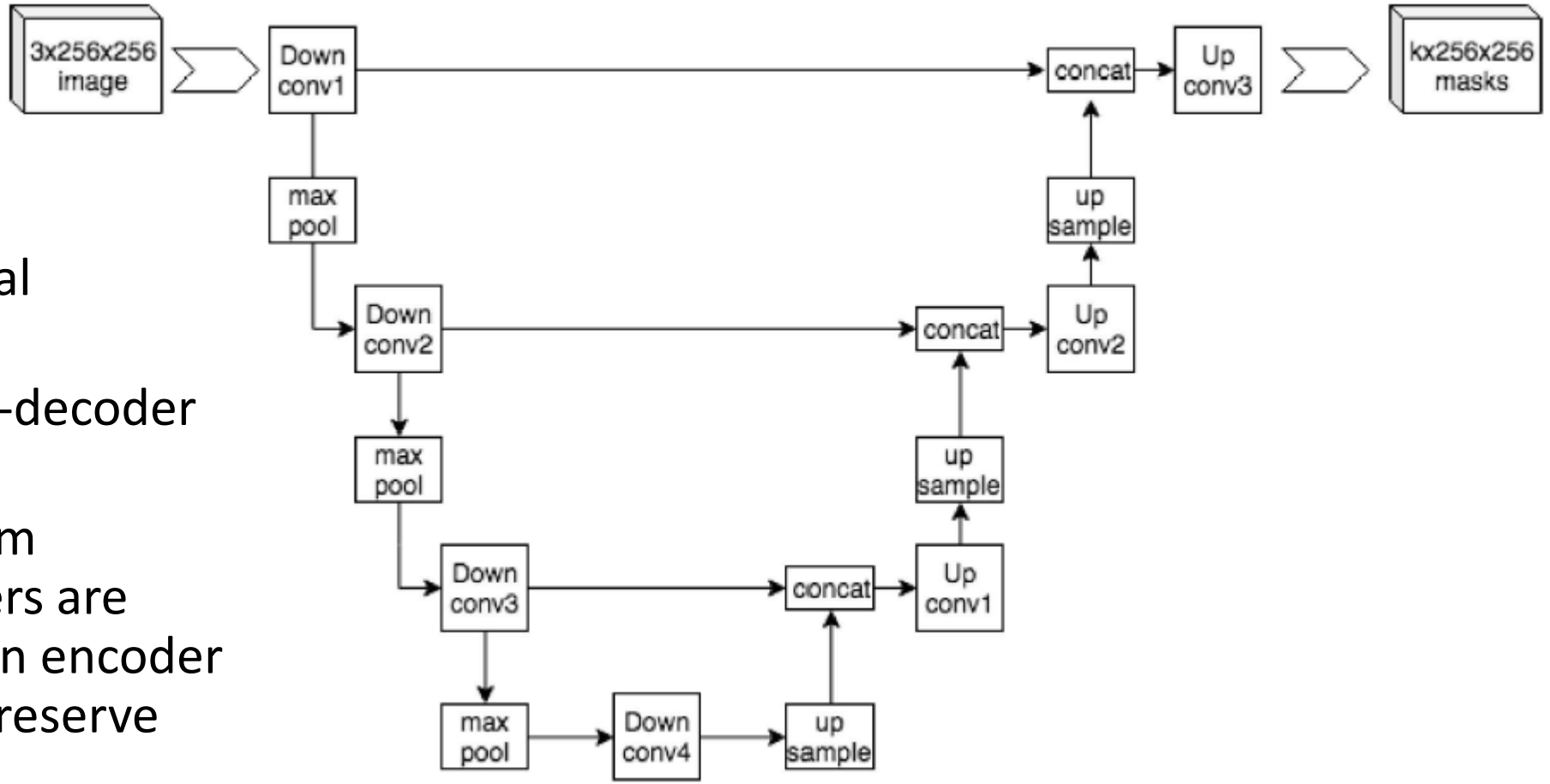
- Semantic segmentation with Deep Learning is performed with a model that takes an image as input and returns the mask of that image that delimits each object or zone into which the image is to be divided.
- A typical semantic segmentation architecture consists of an **encoder** network followed by a **decoder** network:
  1. **Encoders** are typically pre-trained classification networks such as VGG/ResNet. The encodet takes the original image and code the image into a more abstract a with less size with no spatial info
  2. **Decoders** have the task of semantically projecting the low-resolution features learned by the encoder onto a pixel space, at higher resolution, to achieve classification. Takes the output of the encoder and decodes the previous info to produce the output masks.



Some segmentation architectures:

Unet

- Fully convolutional architecture
- Based in encoder-decoder structure
- Feature maps from intermediate layers are conected between encoder and decoder to preserve more spatial info.



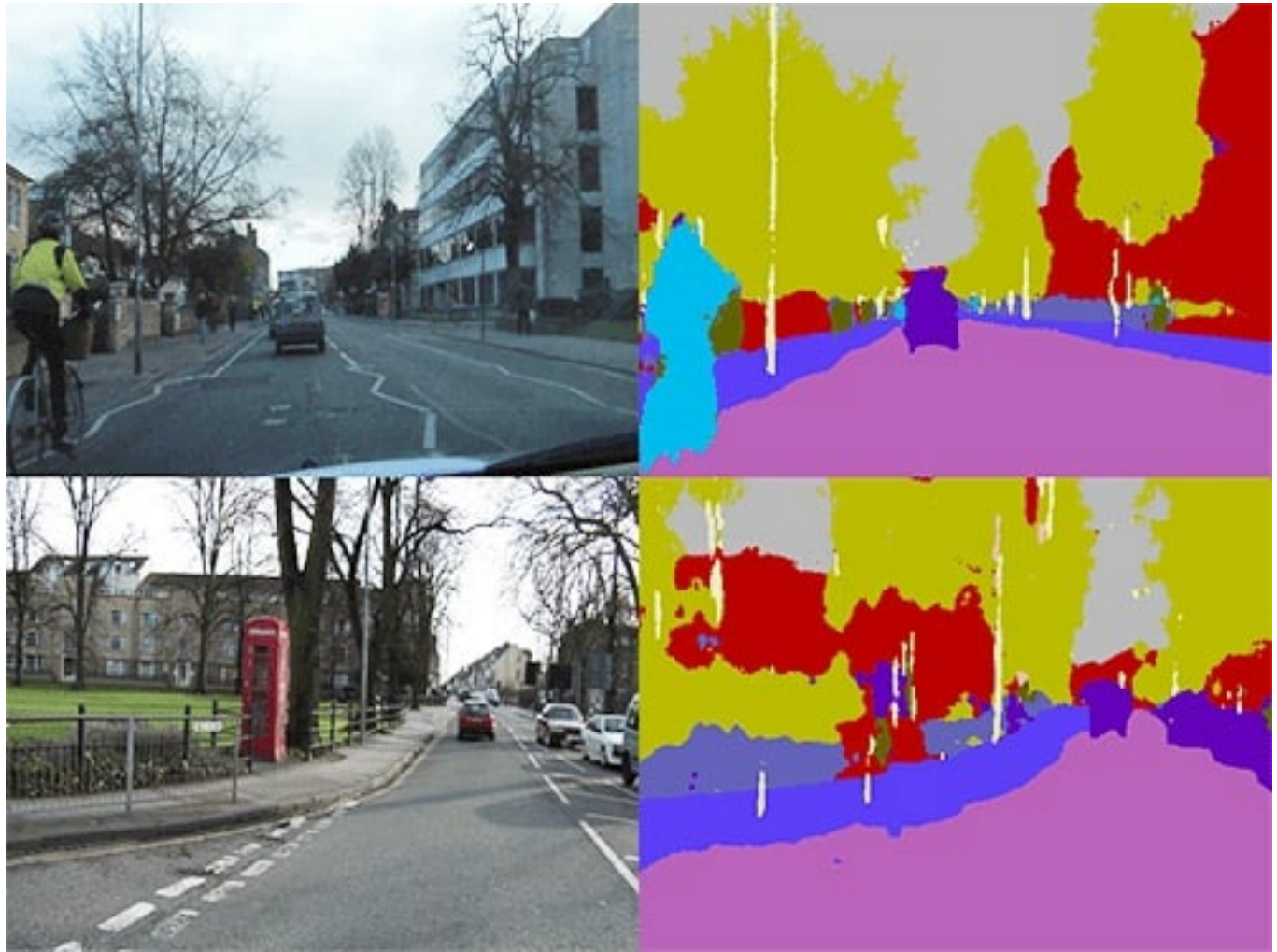
https://arxiv.org/pdf/1505.04597.pdf

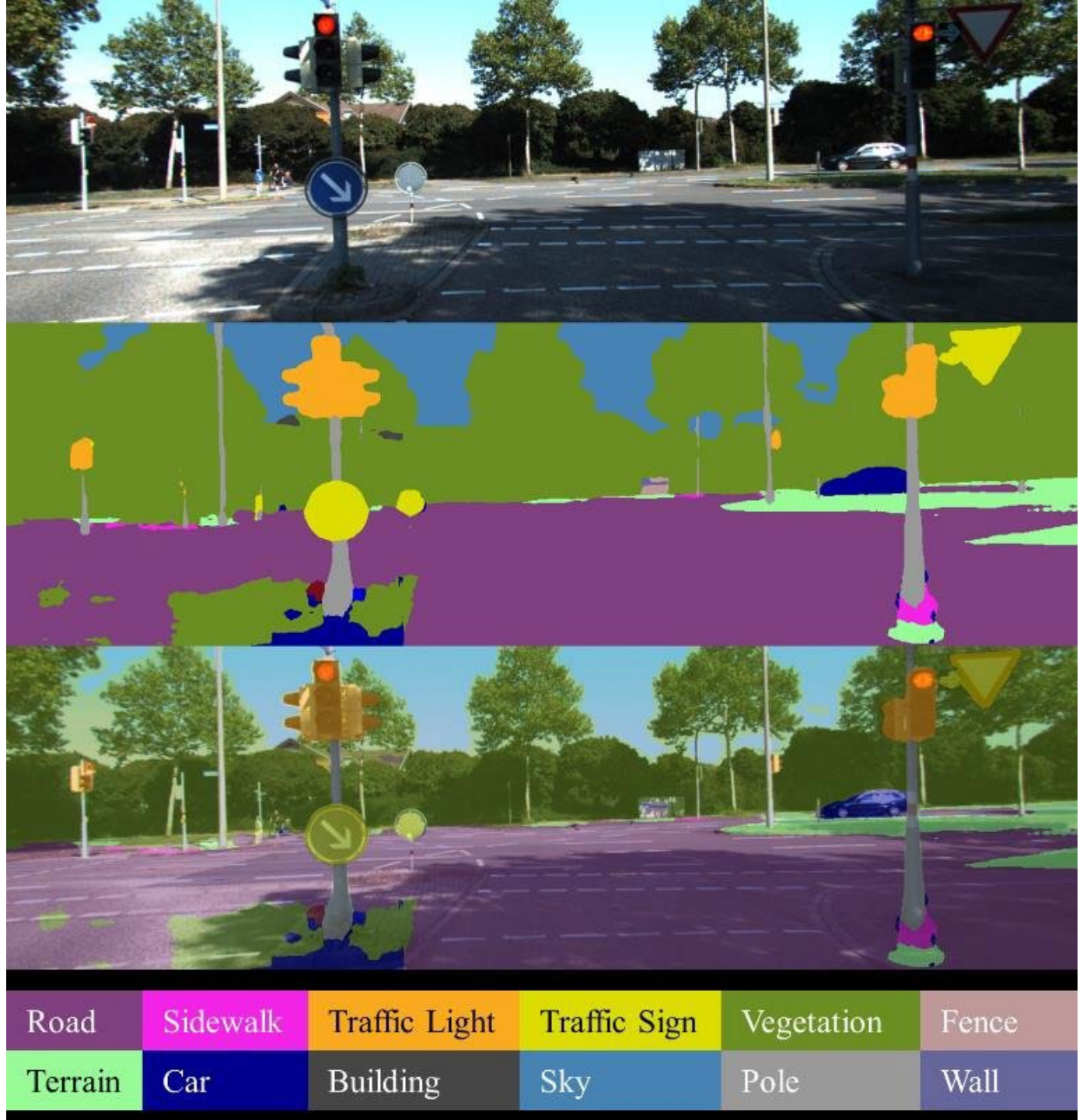
Other architectures are:

- LinkNet
- Feature Pyramid Network (FPN)

Fully Convolutional Network (FCN)-Based Semantic Segmentation

- The original FCN model is able to learn pixel-to-pixel mapping without extracting region proposals. An extension of this model is the FCN network pipeline, which allows existing CNNs to use arbitrary-sized images as input. This made possible because, unlike in a traditional CNN that ends with a fully-connected network with fixed layers, an FCN only has convolutional and pooling layers. The ability to flexibly process images of any size makes FCNs applicable to semantic segmentation tasks.





References:

- *Jason Brownlee*, A Gentle Introduction to Object Recognition With Deep Learning - MachineLearningMastery.com
- Introduction to Object Detection - Machine Learning - HackerEarth Blog
- YOLO: Real Time Object Detection ·pjreddie/darknet Wiki ·GitHub
- GitHub - rbgirshick/py-faster-rcnn: Faster R-CNN (Python implementation) -- see https://github.com/ShaoqingRen/faster_rcnn for the official MATLAB versión
- https://datagen.tech/guides/image-annotation/semantic-segmentation/
- https://www.jeremyjordan.me/semantic-segmentation/