JC3504 Robot Technology

Lecture 9: Computer Vision (3)

Dr Xiao Li xiao.li@abdn.ac.uk

Dr Junfeng Gao Junfeng.gao@abdn.ac.uk

Outline

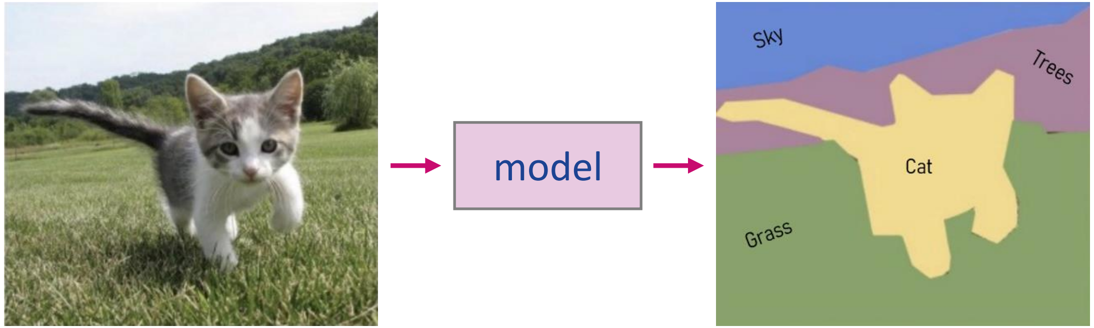
- Image Classification
- Object Detection
- Semantic Segmentation
- Instance Segmentation
- Object Tracking
- Pose Estimation
- Depth Estimation
- Facial Recognition
- …
- Large model solution

Semantic Segmentation

Semantic Segmentation

Semantic segmentation aims to achieve pixel-level classification. Given an image, the semantic segmentation model outputs a segmentation map to annotate the pixel classifications.

The biggest challenge is that pixel classification must be based on pixel context. It's not possible to classify pixels individually.



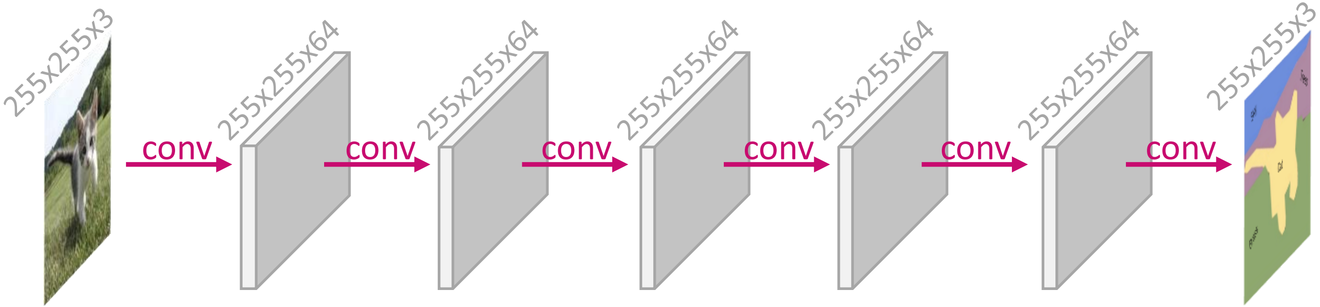
Input Image

Segmentation Map

Semantic Segmentation

The first idea is straightforward. Initially, we need to prepare a segmentation dataset consisting of image pairs, where each pair includes an image and the corresponding segmentation map created by humans.

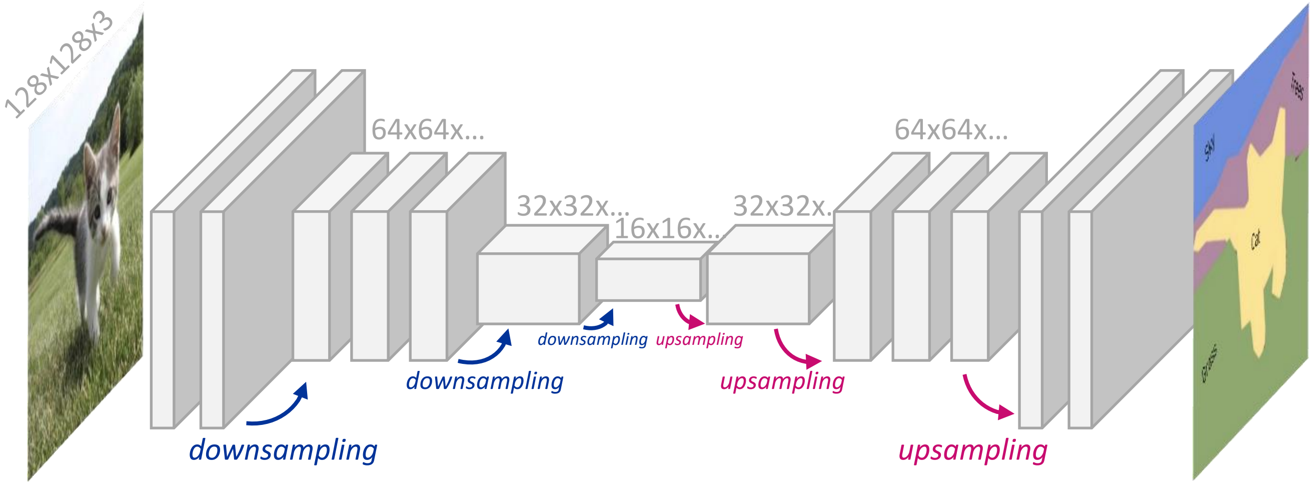
Then, we can design a CNN without any downsampling operators to keep the image size from the begining to the end. This way, we can use the segmentation maps as outputs to train the model.



But this is expensive

Semantic Segmentation

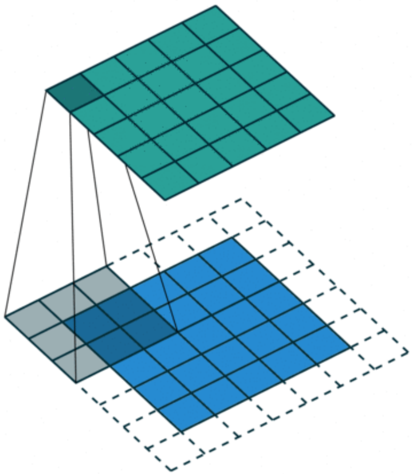
The improved approach involves designing the network as a series of convolutional layers, incorporating both downsampling and upsampling within the network itself.



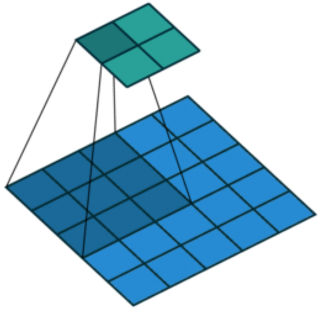
Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015
Noh et al, “Learning Deconvolution Network for Semantic Segmentation”, ICCV 2015

Downsampling

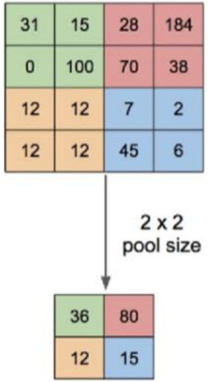
Downsampling reduces the input size. We have introduced ways for downsampling operation.



Conv2d stride=1



Conv2d stride=2 downsampling



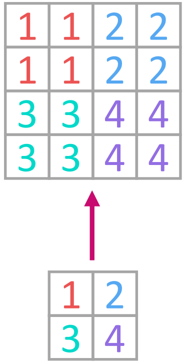
31	15	28	184
0	100	70	38
12	12	7	2
12	12	45	6

36	80
12	15

pooling (stride=2) downsampling

Upsampling

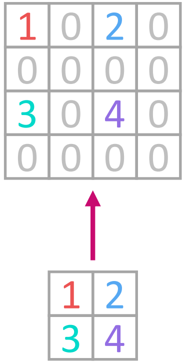
Upsampling, in contrast to downsampling, increases the size of its input. There are some methods that can be used for upsampling.



1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

1	2
3	4

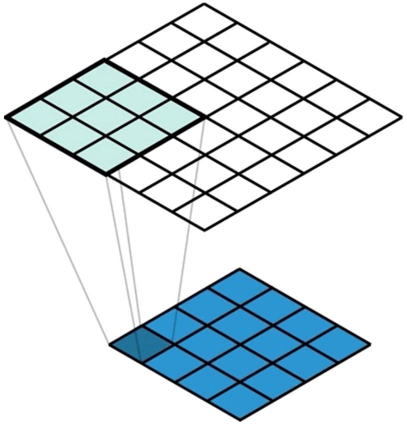
Nearest Neighbours



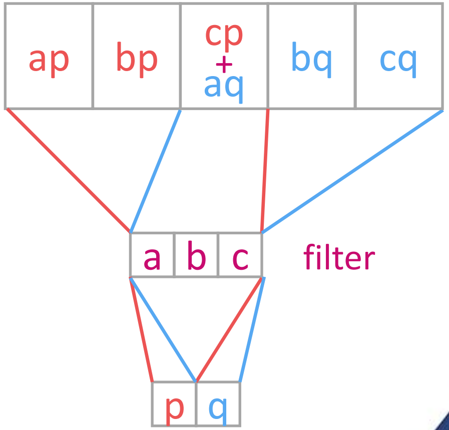
1	0	2	0
0	0	0	0
3	0	4	0
0	0	0	0

1	2
3	4

“Bed of Nails”



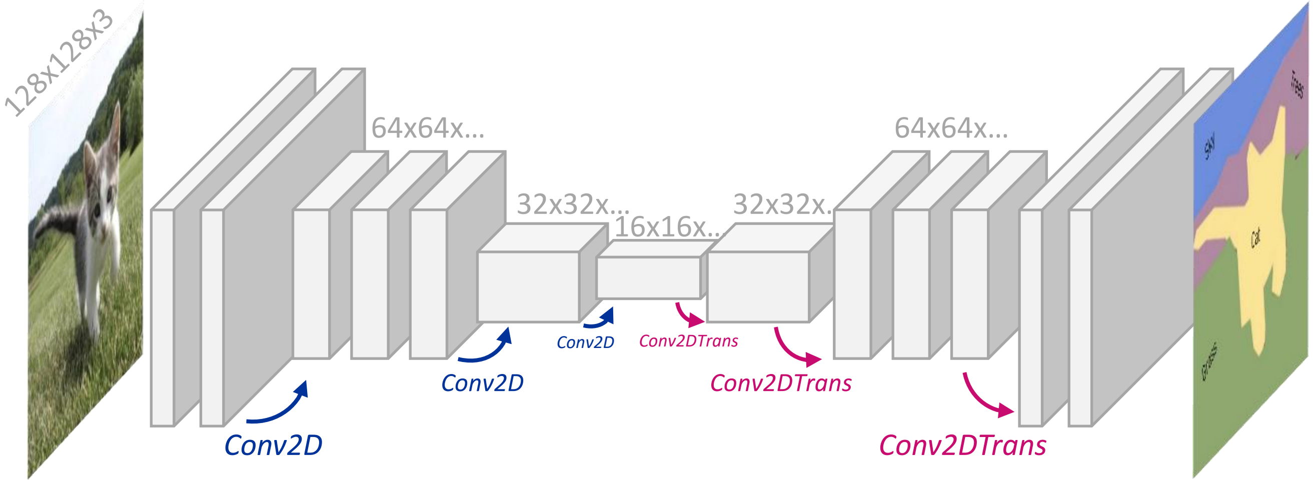
Transposed Convolution



Transposed
Convolution
1D Example

Semantic Segmentation (FCN)

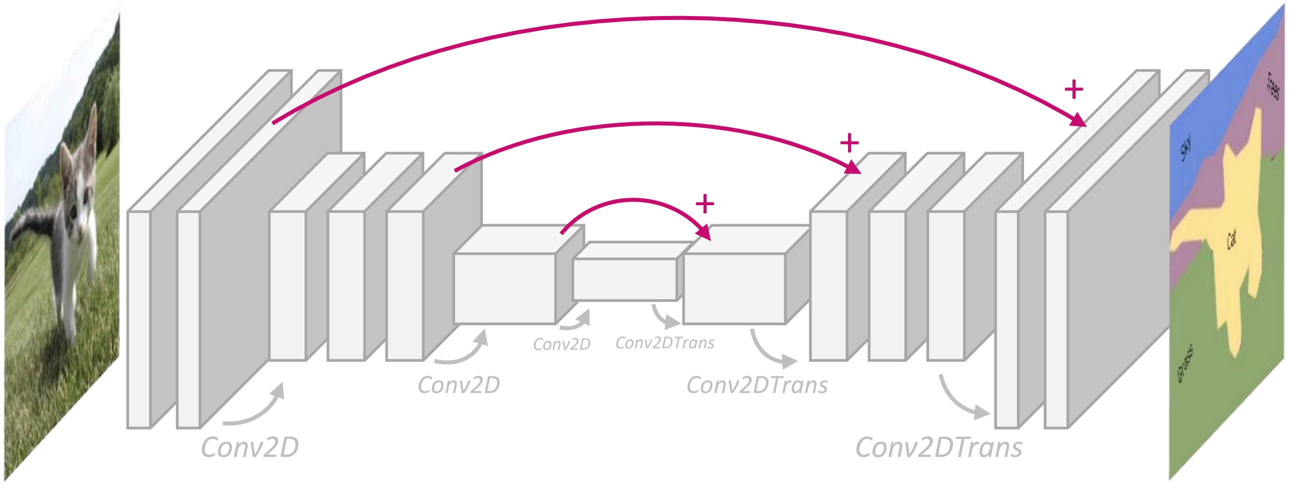
The improved approach involves designing the network as a series of convolutional layers, incorporating both downsampling and upsampling within the network itself.



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 (U-Net)

Another well-known model is U-Net, which adopts residual connections based on a downsampling/upsampling architecture.



Olaf Ronneberger, Philipp Fischer, and Thomas Brox,
"Convolutional Networks for Biomedical Image Segmentation", 2015

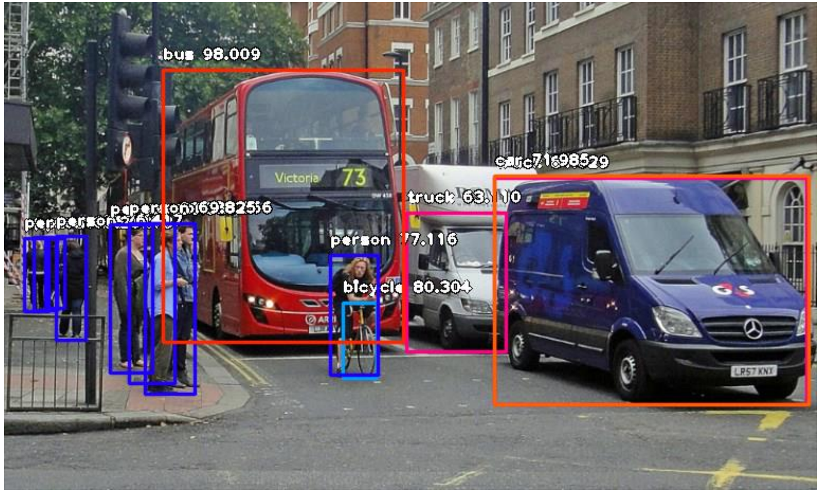
Object Detection

1495 UNIVERSITY OF ABERDEEN

Object Detection

Object detection is the task of identifying and locating objects within an image by classifying them into predefined categories and outlining their positions with bounding boxes.

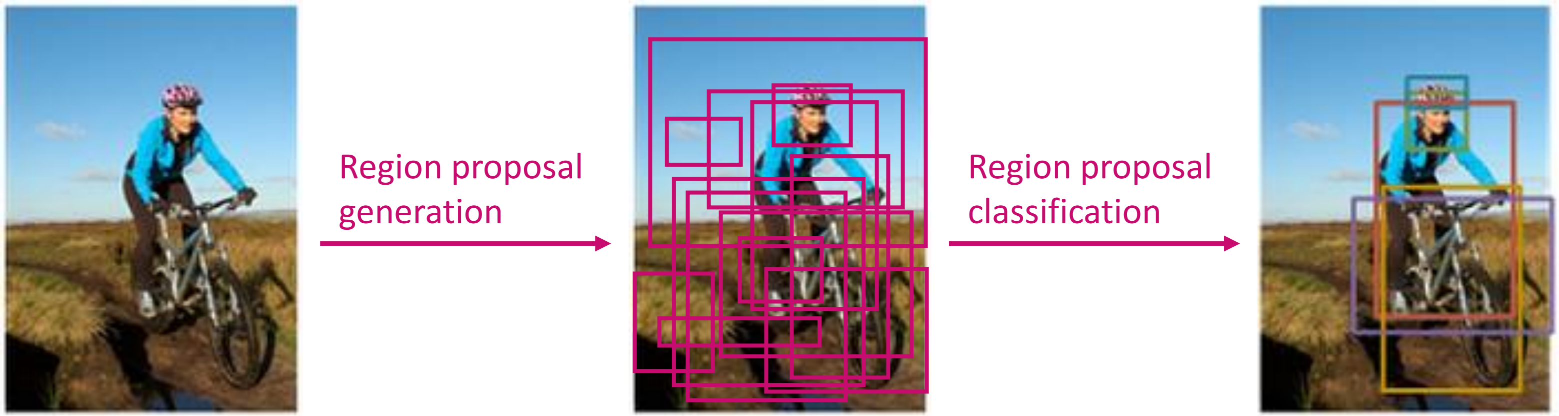
The main challenge of object detection lies in accurately **classifying and localising** objects of various numbers, scales, shapes, and classes within complex contexts.



R-CNN Framework

The R-CNN (Regions with Convolutional Neural Networks) framework detects objects by using a two-step method:

1. selective search to generate **region proposals**, which are then
2. classified and refined using convolutional neural networks and regression models.



Region proposal generation

Region proposal classification

Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014
Ren et al, “Faster r-cnn: Towards real-time object detection with region proposal networks”, NeurIPS 2015

Region Proposals

Region proposals are candidate bounding boxes in an image that are likely to contain objects. They are used by many object detection works.

- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 2000 region proposals

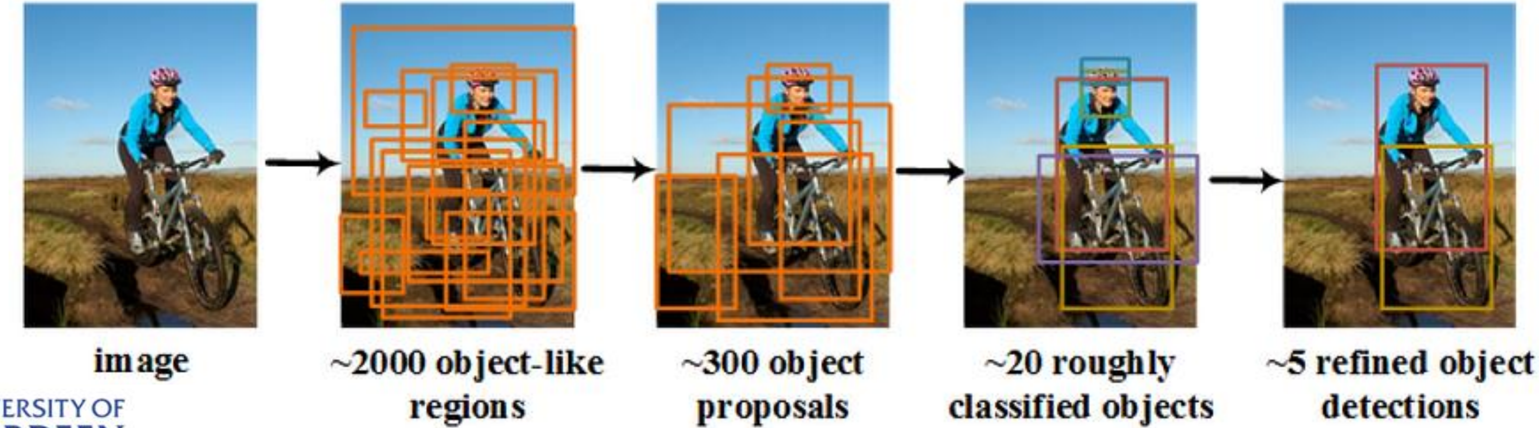
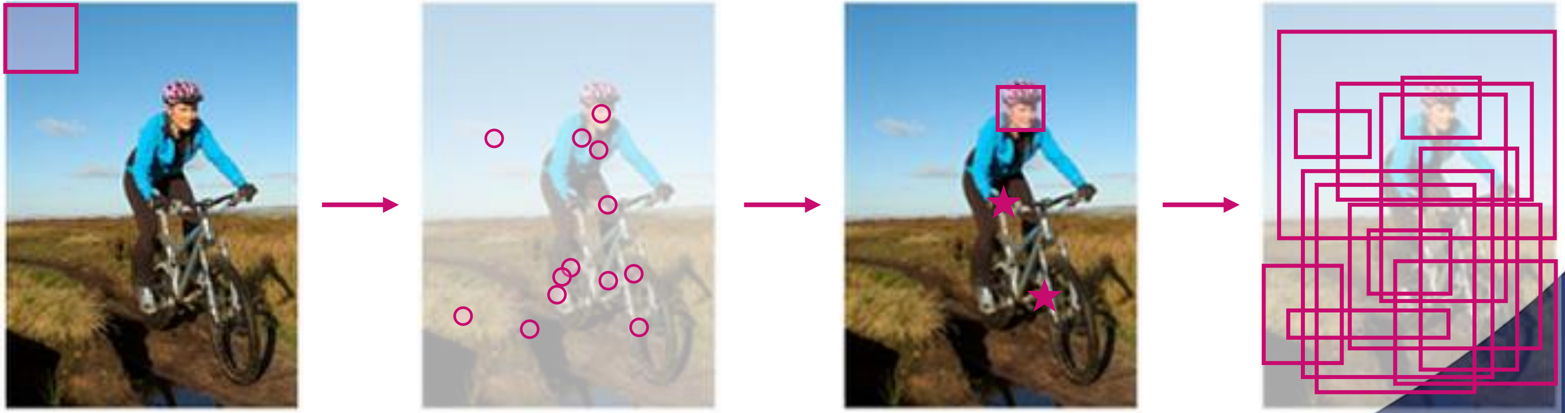


image
~2000 object-like regions
~300 object proposals
~20 roughly classified objects
~5 refined object detections

Region Proposal Network, RPN

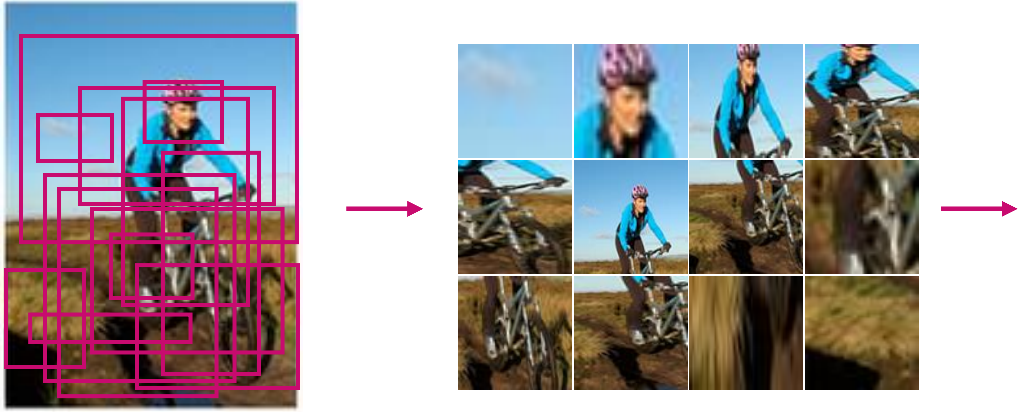
The Region Proposal Network (RPN) is a crucial component of the Faster R-CNN, designed to efficiently generate high-quality region proposals.

- Sliding Window: The RPN scans with a small network (sliding window) centred at every position (i.e., every pixel), to predict if the position contains objects.
- Anchors: For each selected position (the anchor), try bounding boxes of various sizes and ratios and output the best one as a region proposal.



Classification

- The Selective Search algorithm generates approximately 2000 region proposals (also called candidate boxes) in the input image.
- Each region proposal is resized (e.g., scaled) to a fixed size (e.g., 224x224 pixels).
- Each region proposal is classified by a classifier (CNN or SVM classifier).

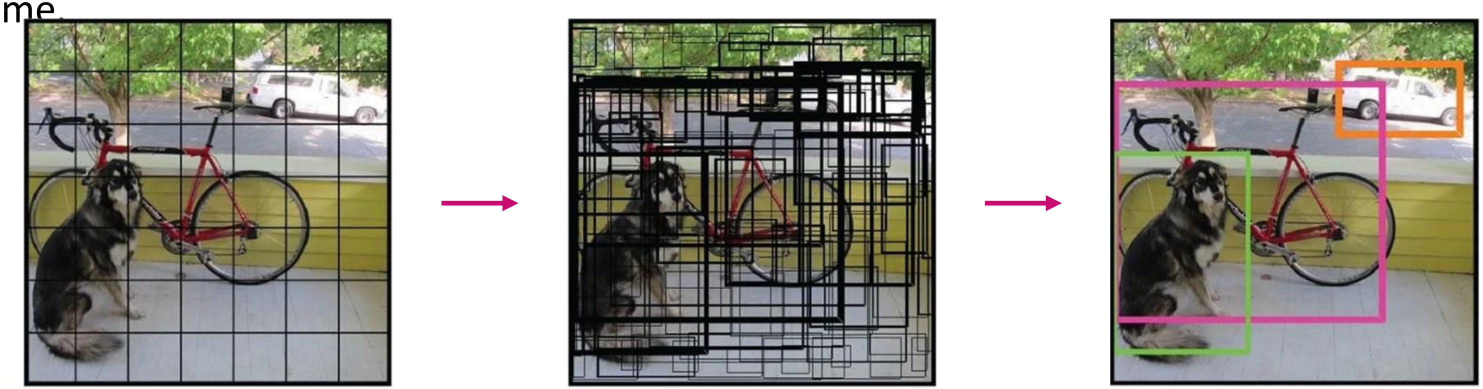


bkg	face	psn face	psn bick
bick psn	psn bkg	bkg bick	bkg bick
bick bkg	bkg bick	bkg	bkg

YOLO Framework

YOLO, or "You Only Look Once", is a ground-breaking object detection algorithm that distinguishes itself from R-CNN by processing images in a single pass, significantly speeding up detection times while maintaining high accuracy.

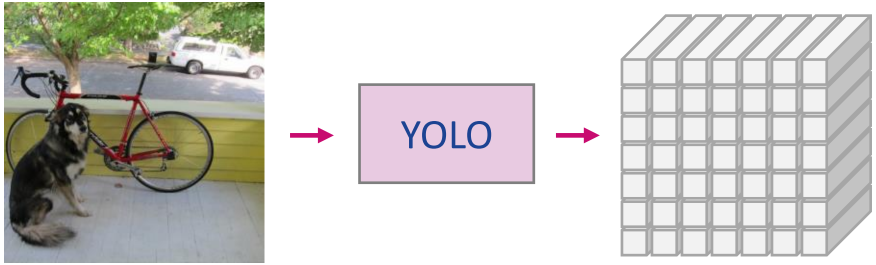
YOLO divides the image into a $S \times S$ grid and predicts bounding boxes and class probabilities for each grid cell in a single forward pass, effectively detecting objects in real-time.



Redmon et al, "You only look once: Unified, real-time object detection.", CVPR 2016

YOLO Framework

Assuming the image is divided into $S \times S$ cells, in the actual prediction process, the entire image is fed into a CNN and, through a series of convolutions and downsampling, ultimately becomes an $S \times S \times H$ tensor i.e. $S \times S$ vectors with the vector length is H.



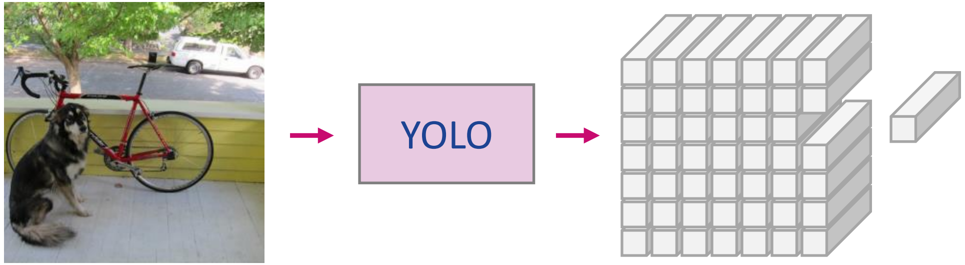
YOLO

Redmon et al, "You only look once: Unified, real-time object detection.", CVPR 2016

YOLO Framework

In the $S \times S$ grid, each vector is responsible for predicting B bounding boxes, including the object class (c), their coordinates (i.e. x, y, width (w) and height (h)) and confidence scores (p), specifically for objects whose centres fall within the cell.

The confidence scores indicate whether the bounding box is in use. If the confidence score approaches 0, the bounding box is not used (don't show). If the score approaches 1, the box is displayed according to its coordinates.



$$= [c, \underbrace{x_1, y_1, w_1, h_1, p_1}_{\text{bounding box 1}}, \underbrace{x_2, y_2, w_2, h_2, p_2}_{\text{bounding box 2}}, \underset{\ldots}{\ldots}]$$

Redmon et al, "You only look once: Unified, real-time object detection.", CVPR 2016

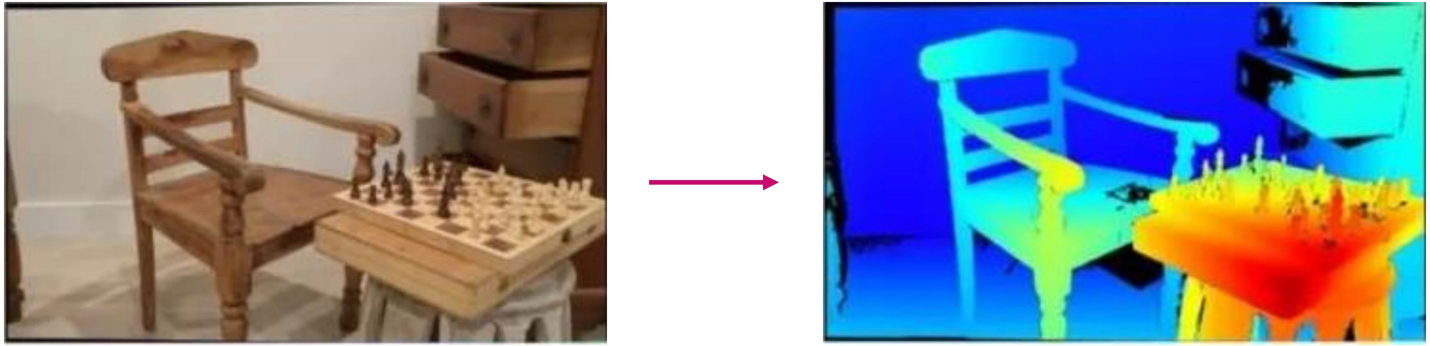
YOLO Framework

<video: 09 - YOLO v7 + SORT Object Tracking.mp4>

Depth Estimation

Depth Estimation

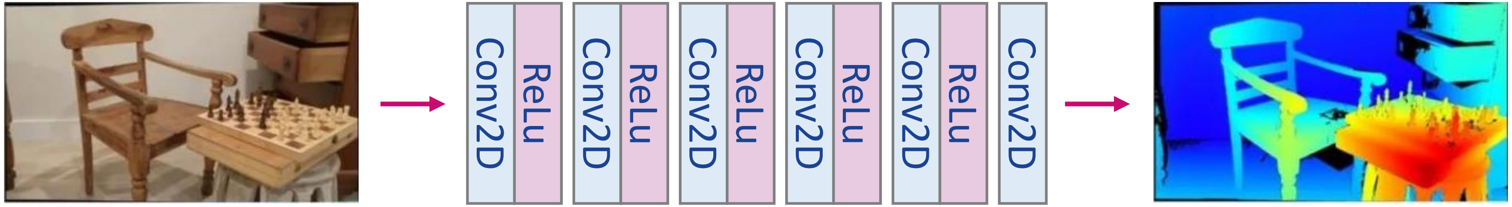
Depth estimation involves predicting the distance of objects from the observer in a scene, typically using images or video data to understand the three-dimensional structure of the environment.

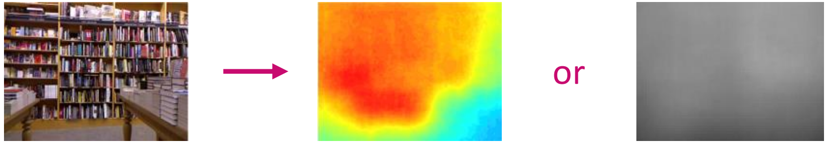


Depth Estimation

The most fundamental approach to implementing depth estimation is to set up an end-to-end CNN network that takes the original image as input and outputs depth predictions, leveraging training datasets that contain both the original images and their corresponding depth maps.

People tried this approach, but the results were not exciting.



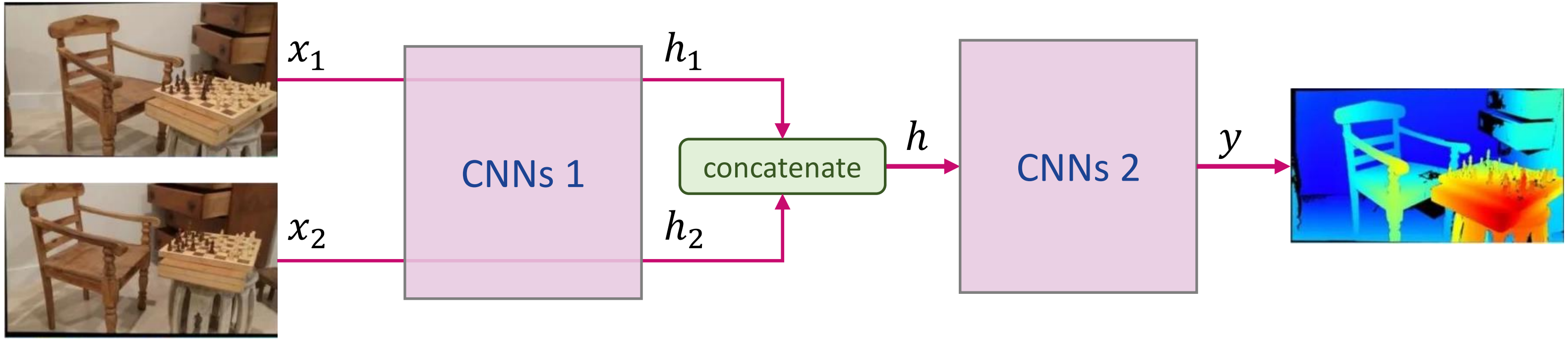


Eigen et al, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network", NeurIPS 2014
Laina et al, "Deeper Depth Prediction with Fully Convolutional Residual Networks" , 3VD 2015

Depth Estimation

A more advanced approach involves the use of images of stereo image pairs (images for left and right eyes).

The left and right images are individually processed by a CNN, after which the results are concatenated and then fed into a second CNN to obtain a depth map.

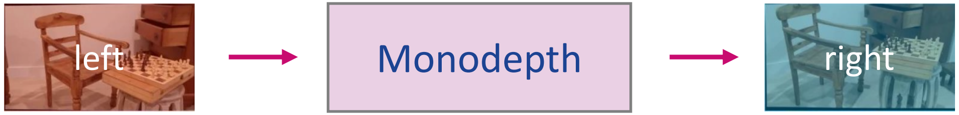


Žbontar et al, "Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches", JMLR 2016

Monodepth Framework

Monodepth is a groundbreaking approach that leverages unsupervised (self-supervised) learning for monocular depth estimation, introducing left-right consistency to enhance the accuracy and reliability of depth predictions from single images. NB: Monodepth only require the left and right images for training, with no depth map is required.

Its fundamental concept is to use the left image to predict the right image.



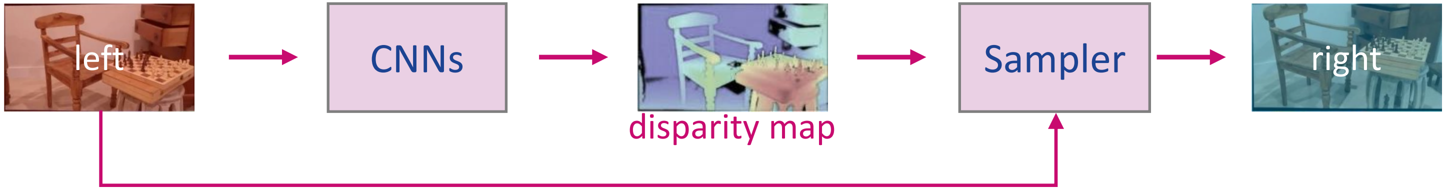
Godard et al, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", CVPR 2017

Monodepth Framework

Monodepth use a CNN to predict the disparity map, which is a 2D tensor that encodes the distance (for each pixel) in horizontal position between corresponding points in stereo images.

Then the network generates the right image with backward mapping using a bilinear sampler.
NB: the sampler is a definite function, not learnable.

So, the generated image can be used to calculate loss with ground true right image.



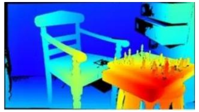
Godard et al, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", CVPR 2017

Monodepth Framework

Given a disparity map, depth maps can be calculated using the lens formula.

$$Depth = \frac{fB}{disparity}$$

- f is the focal length of the camera, usually measured in pixels.
- B is the baseline distance between the two cameras, that is, their physical distance.

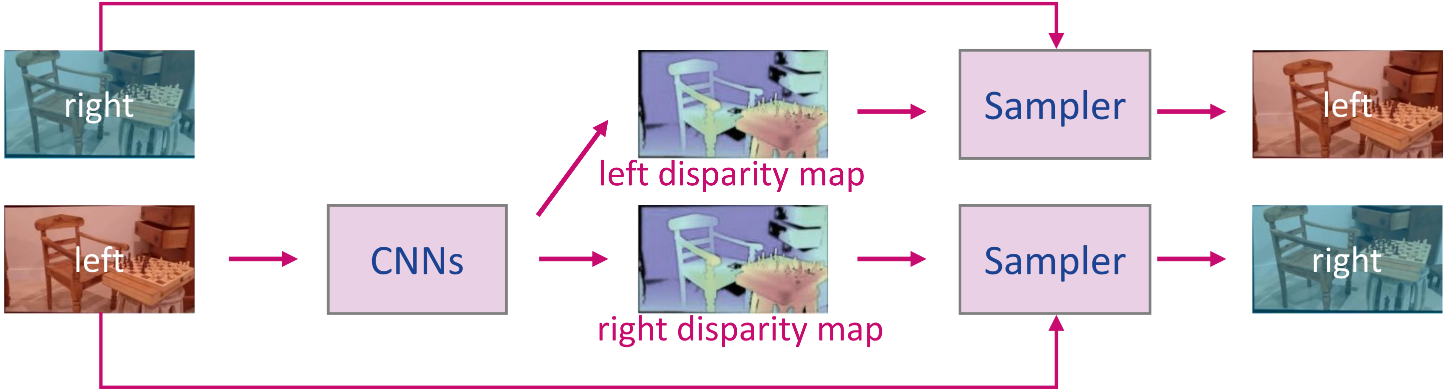


$$= \frac{fB}{\quad}$$

Godard et al, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", CVPR 2017

Monodepth Framework

Finally, Monodepth also proposed a complex model for better performance.



Godard et al, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", CVPR 2017

PROFESSIONAL SNOOKER CLUB
MIGHTYPOUND
YES!! ANY ITEM £1
KFC
99p

Solutions Based on LARGE Models

Solutions Based on Large Models

Equipping robots with vision systems powered by large models offers both significant advantages and certain drawbacks.

- On the positive side, these advanced models can process and interpret complex visual data with high accuracy, enabling robots to navigate environments, recognize objects, and perform tasks with remarkable precision.
- However, the downsides include the high computational resources and energy consumption required to run these large models, potentially limiting their deployment in resource-constrained environments.

Solutions Based on Large Models

<video: Gemini Demo.mp4>

Conclusion

In conclusion, our exploration into the cutting-edge fields of Semantic Segmentation, Object Detection, and Depth Estimation has illuminated their critical roles in advancing machine vision and robotics.

These technologies enable machines to understand and interact with their surroundings in more sophisticated and intuitive ways, paving the path for innovative applications across various industries.