Object Detection Techniques

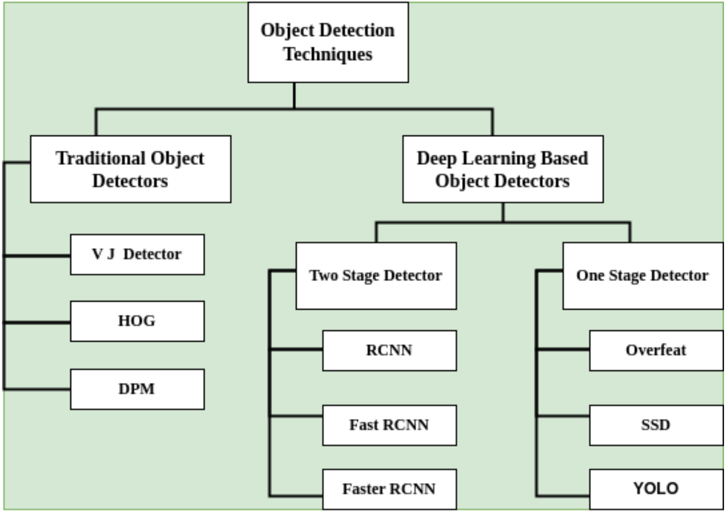

Figure 1: Different Object Detection Methods

Basic Architecture of traditional object detection algorithms

- Due to the limitations of computational resources and datasets traditional object detection methods are still popular.

Figure 2: Steps in traditional object detection [1]

Traditional Object Detectors

- VJ Detector[2]
  - Works by scanning the image or video with a sliding window and applying the classifier at each location.
  - An application to multiscale pedestrian detection is shown that results in nearly real time rates (6fps for 100pixel image) on 640x480 image.
- HOG [3]
  - HOG, or Histogram of Oriented Gradients, is a feature descriptor that is often used to extract features from images.
  - The HOG descriptor focuses on the structure or the shape of an object.
  - Image is partitioned into pixel blocks and in each block we compute a histogram of gradient orientations
- DPM [4]
  - An object detection system based on mixtures of multiscale deformable part models
  - The DPM model starts by borrowing the idea of the HOG detector.
  - A coarse root filter that approximately covers an entire object and higher resolution part filters that cover smaller parts of the object.

Deep Learning based Object Detectors

- Two Stage Detectors
  - First, the model proposes a set of regions of interests by selective search[5] or regional proposal network.
  - Then a classifier only processes the region candidates.Extracts features from each region independently for classification.

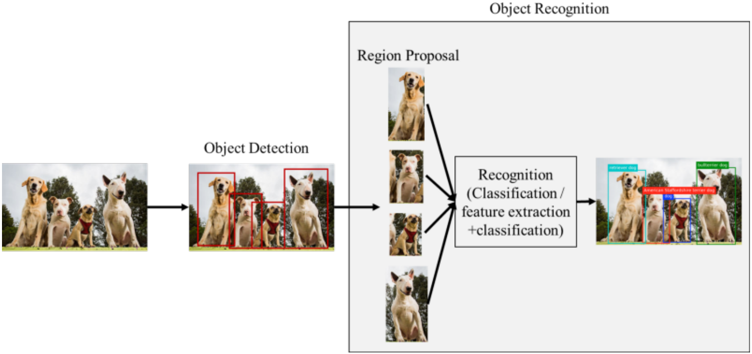

Figure 3: Two Stage Detection

Deep Learning based Object Detectors

- One Stage Detectors
    - Single convolutional network predicts the bounding boxes and the class probabilities for these boxes.

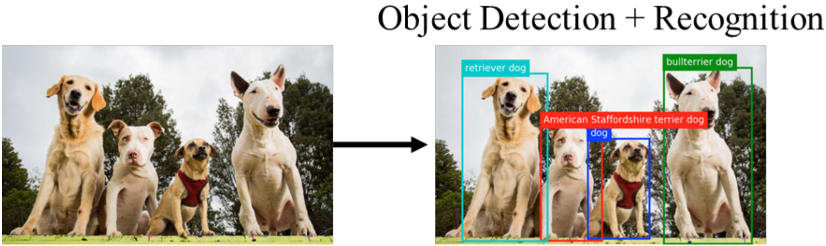

Figure 4: One Stage Detection

Role of Anchors in Object Detectors

- **Anchors Overview:**
  - Anchors are predefined bounding boxes placed across the input image.
- **Functions:**
  - Anchors help generate region proposals and detect objects of interest.
  - They enable object detection at various scales and aspect ratios.
- **Generation Process:**
  - Anchors are created at fixed locations on the input grid.
  - Sizes and aspect ratios are chosen based on dataset analysis.

Anchorbox Predictions

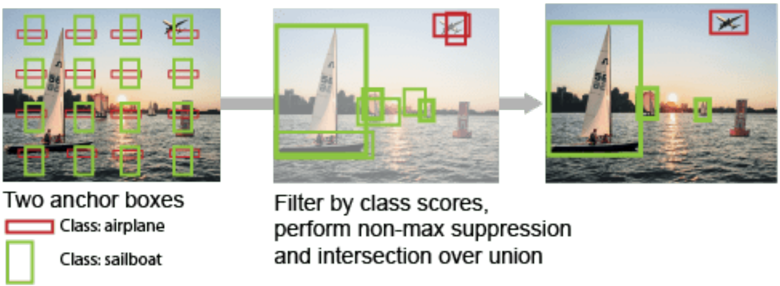

Figure 5: Anchorbox Predictions

Intersection Over Union(IOU)

- IOU
  - A metric used to measure the degree of overlap between two bounding boxes.
  - It calculates the ratio of the area of overlap between the two boxes to the area of their union.
  - IOU $= \frac{\text{Area of Intersection}}{\text{Area of Union}}$

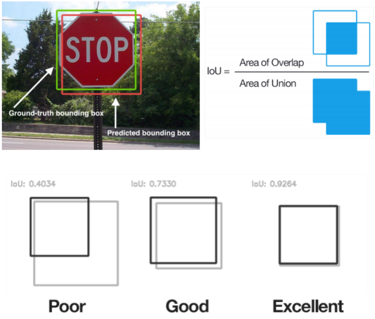

Figure 6: IOU

Non Max Suppression (NMS)

- NMS
  - Used to remove redundancy and select the most accurate proposals
  - Keeps only the proposal with the highest objectness score while suppressing the others.

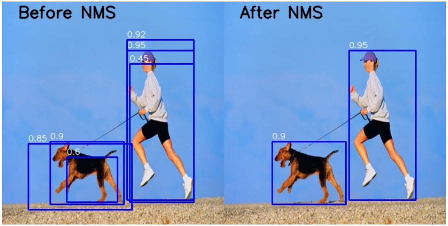

Figure 7: IOU

Deep Learning based Object Detectors

- RCNN [6]
  - An object detection model that uses high-capacity CNNs to bottom-up region proposals in order to localize and segment objects.
  - It uses selective search to identify a number of bounding-box object region candidates ("regions of interest")
  - Extracts features from each region independently for classification.

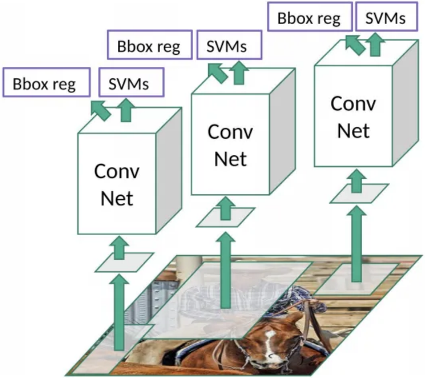

Figure 8: RCNN Network[6]

R-CNN

Problems with R-CNN

- It still takes a huge amount of time to train the network as you would have to classify 2000 region proposals per image.
- It cannot be implemented real time as it takes around 47 seconds for each test image.
- The selective search algorithm is a fixed algorithm. Therefore, no learning is happening at that stage. This could lead to the generation of bad candidate region proposals.

Fast-RCNN

- Fast-RCNN [7]
  - It is similar to the R-CNN algorithm but solved some of the drawbacks of R-CNN
  - Instead of feeding the region proposals to the CNN, the input image is fed to the CNN to generate a convolutional feature map.
  - From the convolutional feature map, we identify the region of proposals and warp them into squares.

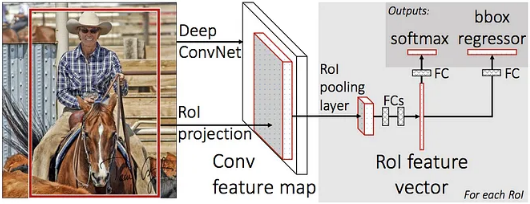

Figure 9: Fast-RCNN Architecture [7]

Fast R-CNN

Problems with Fast R-CNN

- It still uses the Selective Search Algorithm which is slow and a time-consuming process.
- The performance of Fast R-CNN during testing time, including region proposals slows down the algorithm.
- It takes around 2 seconds per image to detect objects, which sometimes does not work properly with large real-life datasets.
- A fast CNN may sacrifice accuracy in order to achieve faster processing times.

Faster-RCNN

- Faster-RCNN [8]
  - Instead of Selective Search algorithm, it uses RPN (Region Proposal Network) to select the best ROIs automatically
  - Similar to Fast R-CNN, the image is provided as an input to a convolutional network which provides a convolutional feature map.

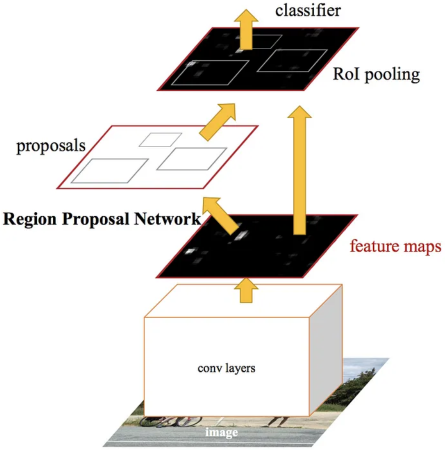

Figure 10: Faster-RCNN Network [8]

Region Proposal Network(RPN)

RPN, Backbone of Faster R-CNN

- RPN has a classifier and a regressor.
- RPNs are designed to efficiently predict region proposals with a wide range of scales and aspect ratios.
- RPNs use anchor boxes that serve as references at multiple scales and aspect ratios.
- Faster R-CNN can struggle with detecting small objects, particularly when they are surrounded by other objects or occluded.

Faster R-CNN

Problems with Faster R-CNN

- Faster-RCNN has some shortcomings such as it has not reached to the real-time detection.
- Faster R-CNN requires a lot of computation and can be slow, especially when processing large datasets.
- Faster R-CNN has many hyperparameters that need to be carefully tuned to achieve good performance. Optimizing these hyperparameters can be time-consuming and difficult.
- Faster R-CNN can struggle with detecting small objects, particularly when they are surrounded by other objects or occluded.

Losses for RCNN

- **Region Proposal Network (RPN) Loss:**
  - Classification Loss: Binary cross-entropy (logistic loss)
  - Regression Loss: Smooth L1 loss
- **Fast R-CNN Loss:**
  - Classification Loss: Cross-entropy loss
  - Regression Loss: Smooth L1 loss
- **Faster R-CNN Loss:**
  - Localization Loss: Combination of classification and regression components

One Stage Detectors

Named based on number of stages involved in the detection process

- Overfeat [9]
  - OverFeat was one of the first modern one-stage object detector based on deep networks.
  - Ability to detect objects at multiple scales and aspect ratios.
  - Use of a single neural network for both feature extraction and object detection allows for a more efficient and accurate detection process.
  - OverFeat algorithm relies on a fixed set of scales and aspect ratios, which may not be optimal for all types of images and objects.
- SSD [10]
  - SSD is designed for object detection in real-time.
  - The SSD algorithm uses a multi-scale feature pyramid to detect objects at different scales and resolutions.
  - SSD has a 10-20% lower AP than two stage detectors

YOLO - You Look Only Once

- YOLO [11] is a popular object detection algorithm that has revolutionized the field of computer vision.
- YOLO uses the fixed grid detector, which makes the technique fast.
- A single neural network is applied to the whole image to detect objects in this technique.
- The whole image is divided into fixed regions, from each region, the probability and bounding box of the object is calculated.

YOLO Architecture

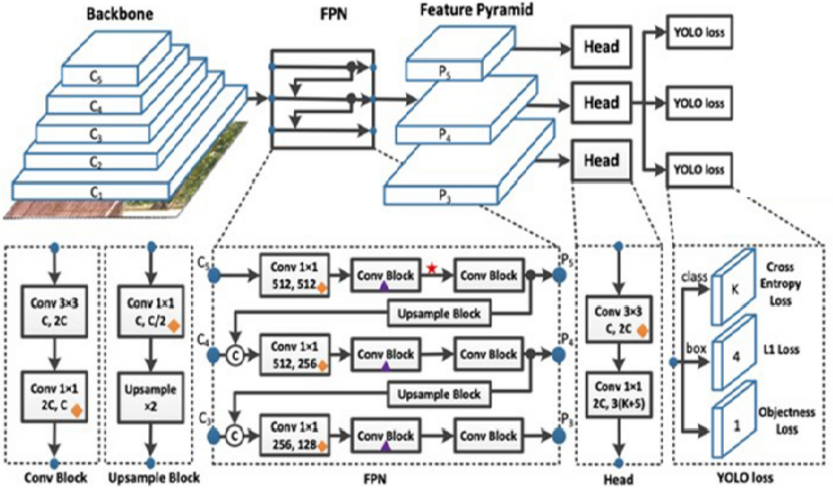

Figure 11: YOLO Architecture

YOLO Predictions

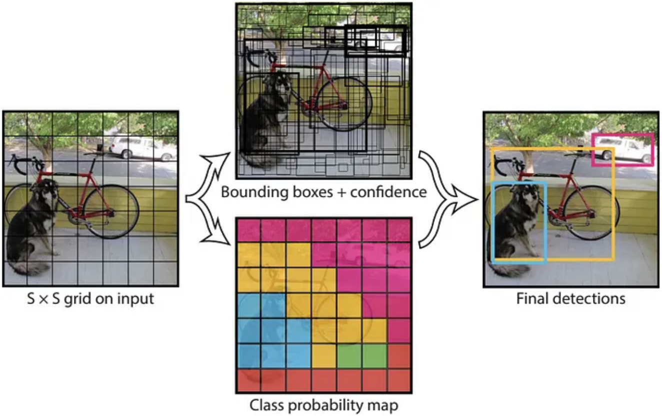

Figure 12: YOLO Model

YOLO - You Look Only Once

- Scholars have published several YOLO subsequent versions described as YOLO V2, YOLO V3, YOLO V4 [12], YOLO V5, V7 [13] YOLO V8 and YOLO V9.
- It was written and is maintained in a framework called Darknet.
- YOLOv5 is the first of the YOLO models to be written in the PyTorch framework and it is much more lightweight and easy to use.

YOLO : Advantages and Disadvantages

Advantages

- Fast: YOLO is a real-time object detection model that can detect objects in an image or video stream in just one pass.
- High Accuracy: YOLO is known for its high accuracy in detecting objects of different sizes and shapes in an image or video.
- Generalizability: YOLO can be used to detect objects in a wide range of applications, including self-driving cars, surveillance systems, and medical imaging.

Disadvantages

- Small Objects: While YOLO can detect small objects, it may not be as accurate as other object detection models in detecting extremely small objects.
- Occlusion: YOLO may struggle to detect objects that are partially occluded by other objects in the image.
- Accuracy in Complex Scenes: YOLO may struggle to accurately detect objects in complex scenes that contain many objects or have a cluttered background

BiFPN(Weighted Bi-directional Feature Pyramid Network)

- A type of feature pyramid network that helps with easy and fast multi-scale feature fusion.
- Allow information to flow both top-down and bottom-up while using regular and efficient connections.
- The BiFPN is designed to treat input features with varying resolutions equally.
- The network can be trained on images of varying resolutions, allowing it to adapt to different tasks and scenarios.
- YOLOv5 is the first of the YOLO models to be written in the PyTorch framework and it is much more lightweight and easy to use.

BiFPN - EfficientDet Architecture [14]

- Introduced by Tan et al. in EfficientDet: Scalable and Efficient Object Detection

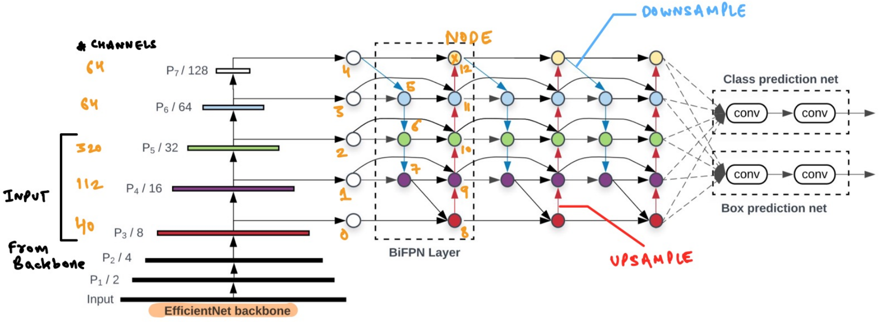

Figure 13: EfficientDet architecture [14]

BiFPN - Applications

- The BiFPN is an efficient and effective method for feature extraction and fusion in computer vision and machine learning tasks.
- Its design allows for easy and fast multi-scale feature fusion, making it ideal for tasks like object detection and semantic segmentation.
- By using a feature pyramid network like the BiFPN, it is possible to detect objects at different scales and resolutions, allowing for more comprehensive detection and improved performance.
- It can also been used in semantic segmentation tasks, where it can improve the accuracy of segmentation by better fusing features at multiple scales.

References I

[1] Youzi, Xiao Tian, Zhiqiang Yu, Jiachen Zhang, Yinshu Liu, Shuai Du, Shaoyi Lan, Xuguang. (2020). A review of object detection based on deep learning. Multimedia Tools and Applications. 79. 10.1007/s11042-020-08976-6.

[2] Piotr Dollar, Serge Belongie, and Pietro Perona. The Fastest Pedestrian Detector in the West. In Frédéric Labrosse, Reyer Zwiggelaar, Yonghuai Liu, and Bernie Tiddeman, editors, Proceedings of the British Machine Vision Conference, pages 68.1-68.11. BMVA Press, September 2010. doi:10.5244/C.24.68.

[3] Dalal, Navneet Triggs, Bill. (2005). Histograms of Oriented Gradients for Human Detection. Comput. Vision Pattern Recognit.. 1. 886-893. 10.1109/CVPR.2005.177.

[4] Felzenszwalb, Pedro Girshick, Ross Mcallester, David Ramanan, Deva. (2010). Object Detection with Discriminatively Trained Part-Based Models. IEEE transactions on pattern analysis and machine intelligence. 32. 1627-45. 10.1109/TPAMI.2009.167.

[5] Uijlings, Jasper Sande, K. Gevers, T. Smeulders, A.W.M.. (2013). Selective Search for Object Recognition. International Journal of Computer Vision. 104. 154-171. 10.1007/s11263-013-0620-5.

References II

[6] Girshick, Ross Donahue, Jeff Darrell, Trevor Malik, Jitendra. (2013). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 10.1109/CVPR.2014.81.

[7] Girshick, Ross. (2015). Fast r-cnn. 10.1109/ICCV.2015.169.

[8] Ren, Shaoqing He, Kaiming Girshick, Ross Sun, Jian. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39. 10.1109/TPAMI.2016.2577031.

[9] Sermanet, Pierre Eigen, David Zhang, Xiang Mathieu, Michael Fergus, Rob Lecun, Yann. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. International Conference on Learning Representations (ICLR) (Banff).

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016.

[11] Redmon, Joseph Divvala, Santosh Girshick, Ross Farhadi, Ali. (2016). You Only Look Once: Unified, Real-Time Object Detection. 779-788. 10.1109/CVPR.2016.91.

References III

[12] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection.

[13] Horvat, Marko Jelečević, Ljudevit Gledec, Gordan. (2022). A comparative study of YOLOv5 models performance for image localization and classification.

[14] Tan, M., Pang, R., Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10781-10790).

THANK YOU