# Task: 08

## YOLO

YOU ONLY LOOK ONCE -- V4

**Architecture:**

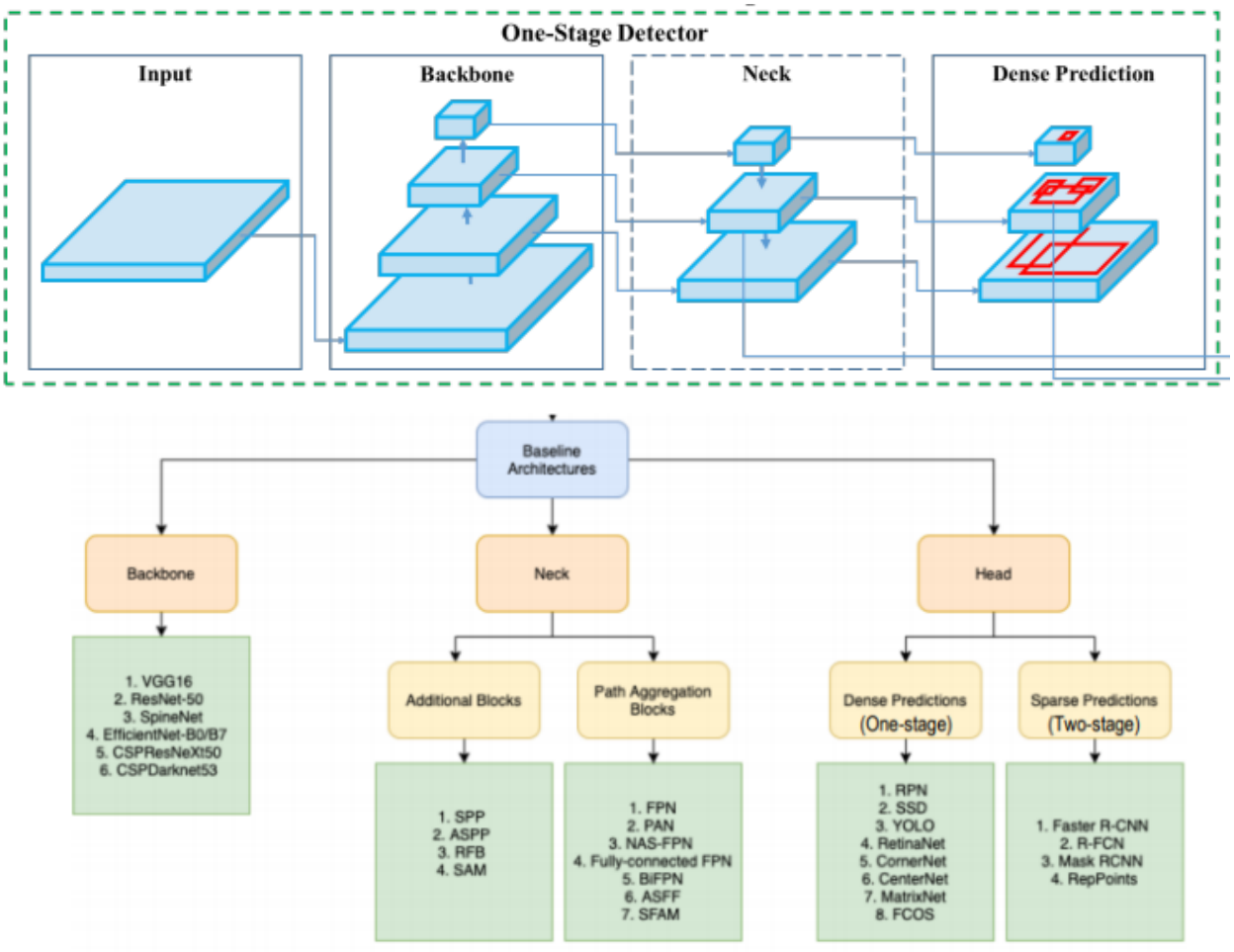



*Figure 16. Diagrams of the most advanced innovation ideas applied by the author to each part of the YOLOv4 architecture.*

YOLOv4 consists of:

BACKBONE: - CSPDarknet53

NECK: - Additional block --> Spatial Pyramid Pooling (SPP)

Path-Aggregation blocks --> PANet

HEAD: - YOLOv3

Backbone – CSPDarknet53 :

The backbone (feature extractor) of the YOLOv4 model was considered by the authors
among 3 options: CSPResNext53, CSPDarknet53 and EfficientNet-B3, the most advanced
convolutional network at that time. Based on theoretical justification and lots of experiments,
CSP Darknet53 neural network was determined to be the most optimal model.

The CSPResNext50 and the CSPDarknet53 (CSP stands for Cross Stage Partial) are both derived from the DenseNet architecture which uses the previous input and concatenates it with the current input before moving into the dense layer (Huang, et al., 2018). DenseNet was designed to connect layers in a very deep neural network with the aim of alleviating vanishing gradient problems (as ResNet).

### Neck (feature aggregation):

Object detectors composed of a backbone in feature extraction and a head for object detection. And to detect objects at different scales, a hierarchy structure is produced with the head probing feature maps at different spatial resolutions. To enrich the information that feeds into the head, neighbouring feature maps coming from the bottom-up stream and the top-down stream are added together element-wise or concatenated before feeding into the head.

YOLOv4 adds a SPP block after CSPDarknet53 to increase the receptive field and separate out the most important features from the backbone, thus improving the model accuracy with negligible increase of inference time. The diagram below demonstrates how SPP is integrated into YOLO

### Head – YOLOv3

In the case of a one-stage detector, the function of the head is to perform dense predictions. The dense prediction is the final prediction composed of a vector containing the predicted bounding box coordinates (centre, height, width), the prediction confidence score, and the probability classes. YOLOv4 deploys the identical head as YOLOv3 for detection with the anchor-based detection steps, and three levels of detection granularity.

## Techniques used to improve object detection:

1. Bag of Freebies (BoF) for backbone:
   CutMix and Mosaic data augmentation, DropBlock regularization,Class label smoothing

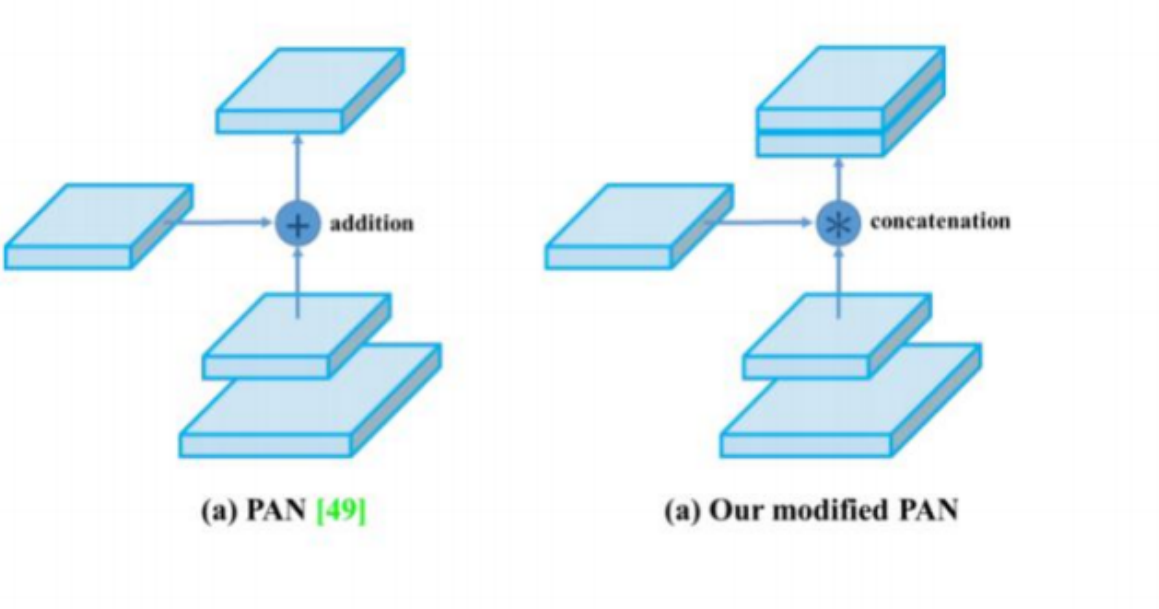
2. Bag of Specials (BoS) for backbone:
   Mish activation, Cross-stage partial connections (CSP), Multi-input weighted residual connections (MiWRC).

3. Bag of Freebies (BoF) for detector:
   CIoU-loss, CmBN, Drop Block regularization, Mosaic data augmentation, Self-Adversarial Training, Eliminate grid sensitivity, Using multiple anchors for a single ground truth, Cosine annealing scheduler, Optimal hyper-parameters, Random training shapes.

4. Bag of Specials (BoS) for detector:
   Mish activation, SPP-block, SAM-block, PAN path-aggregation block, DIoU-NMS

## Other Improvements:

YOLOv4 uses PANet for the feature aggregation of the network. However, instead of adding neighbour layers together, features maps are concatenated together in YOLOv4. Conclusion: YOLO v4 achieves state-of-the-art results (43.5% AP) for real-time objects.




## Conclusion:

YOLO v4 achieves state-of-the-art results (43.5% AP) for real-time object detection and is able to run at a speed of 65 FPS on a V100 GPU.