[Week 12] Object Detection

SWE3032-41 Artificial Intelligence Project

Mann Soo Hong[1], Soon Cheol Noh[2]
msjo91@skku.edu[1], yhnuhb27@skku.edu[2]
SKKU Information & Intelligence Lab
10/05/2021

Confusing Terminology

- Localization, Segmentation, Detection, Recognition?

Name	Task	Output
Localization	Finding the positions	Bounding box, Coordinate
Segmentation	Finding the boundaries	Segment
Detection	Localization (+ Classification)	Bounding box (+ Class)
Recognition	Detection	Bounding box + Class

Confusing Terminology

- Image, Scene, Object, Semantic, Instance, Scene Segmentation?

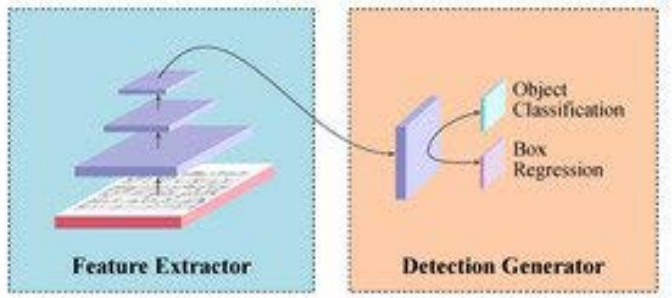
Segmentation	Task
Image	Splitting an image into segments
Scene	Splitting a scene into object components
Object	Finding the boundaries of objects
Semantic	Finding the boundaries and differentiating objects by class
Instance	Finding the boundaries and differentiating each objects

Object Detection is Basically…

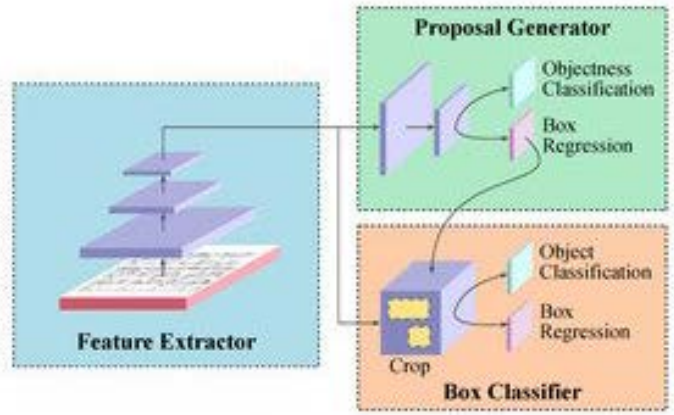
$$\boldsymbol{Object\ Detection = Object\ Localization + Classification}$$

- This means you can:
  - Find possible regions of interest (ROI) then classify (2-stage)
  - Do them simultaneously (1-stage)
  - Encode then decode (U-Net)

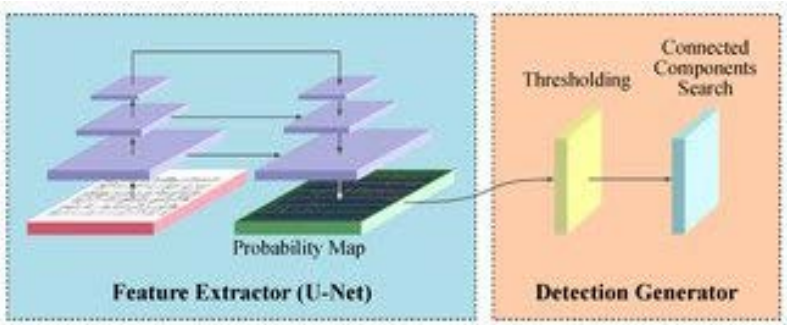
Object Detection Architectures



(a) Basic architecture of a one-stage detector.



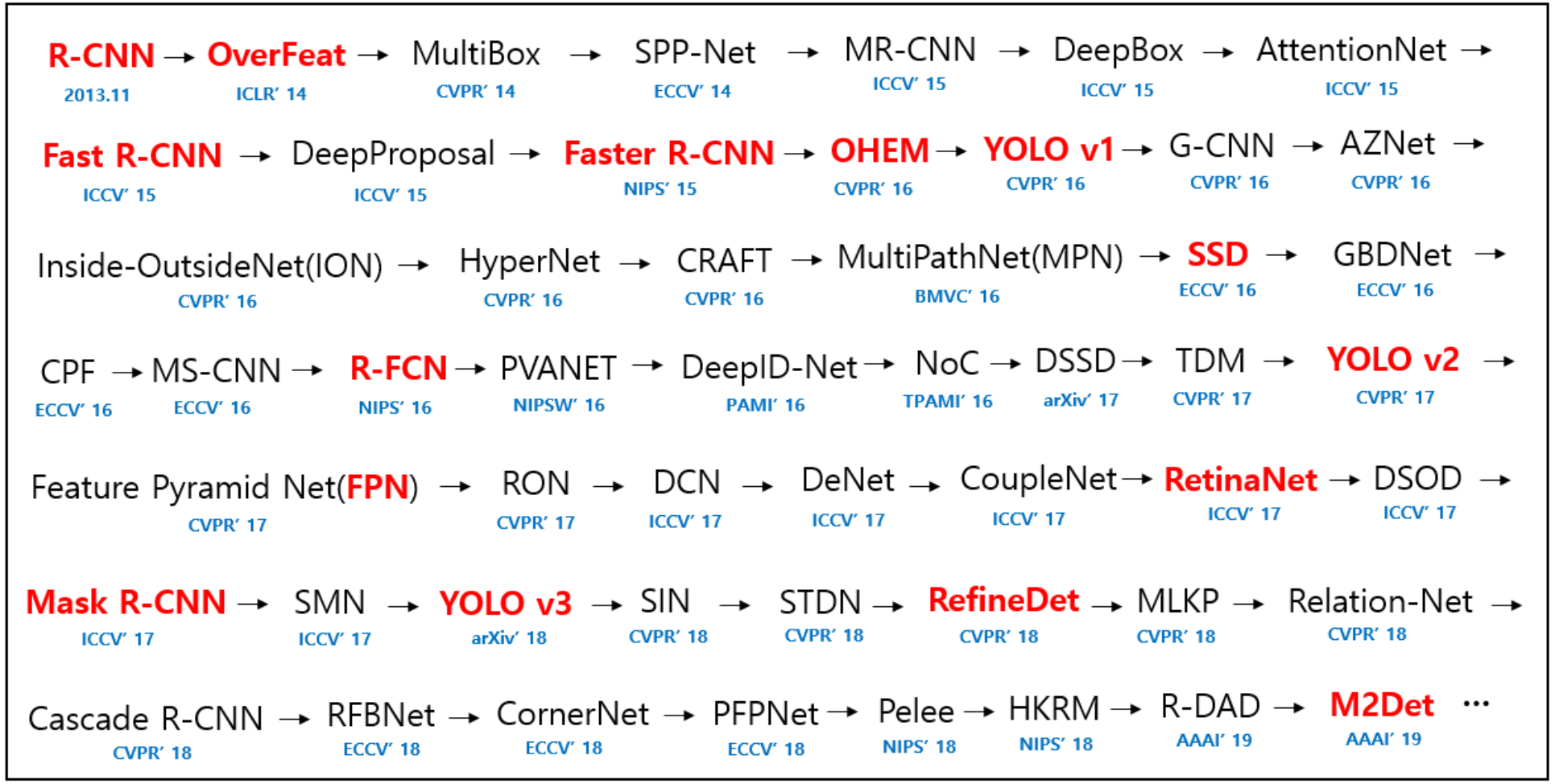
(b) Basic architecture of a two-stage detector.



(c) Basic architecture of the U-Net detector.

A. Pacha, et al. 2018. “A Baseline for General Music Object Detection with Deep Learning” Appl. Sci. 8, no. 9: 1488. https://doi.org/10.3390/app8091488

Development of Object Detection



H. Lee. 2018. "deep learning object detection". GitHub repository, https://github.com/hoya012/deep_learning_object_detection

YOLOs

- **Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. (2016) *You Only Look Once: Unified, Real-Time Object Detection*. In CVPR, 2016.**
- Joseph Redmon and Ali Farhadi. (2017) *YOLO9000: Better, Faster, Stronger*. In CVPR, 2017.
- Joseph Redmon and Ali Farhadi. (2018) *YOLOv3: An Incremental Improvement*.
- Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. (2020) *YOLOv4: Optimal Speed and Accuracy of Object Detection*.
- Glenn Jocher et al. (2020) *YOLOv5*.

YOLO in Short

You Only Look Once: Unified, Real-Time Object Detection

- Look Once: Scans the whole image once
- Unified: Uses a single neural network end-to-end
- Real-Time: Fast object detection of 45 FPS

Detection System

YOLOv1

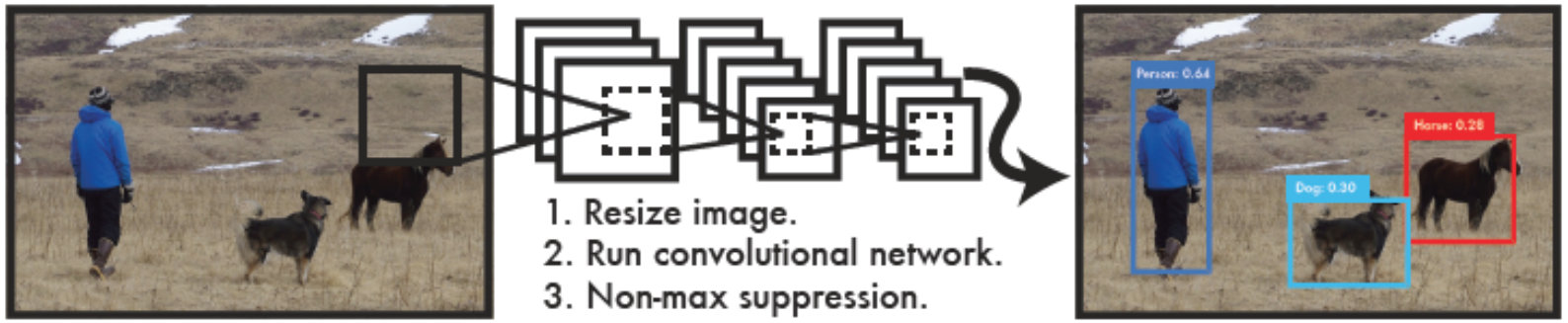


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

- No ROI extraction
- Straight to bounding box and classification

J. Redmon, et al. 2016. "You Only Look Once: Unified, Real-Time Object Detection". In 2016 CVPR.

Detection System

YOLOv1

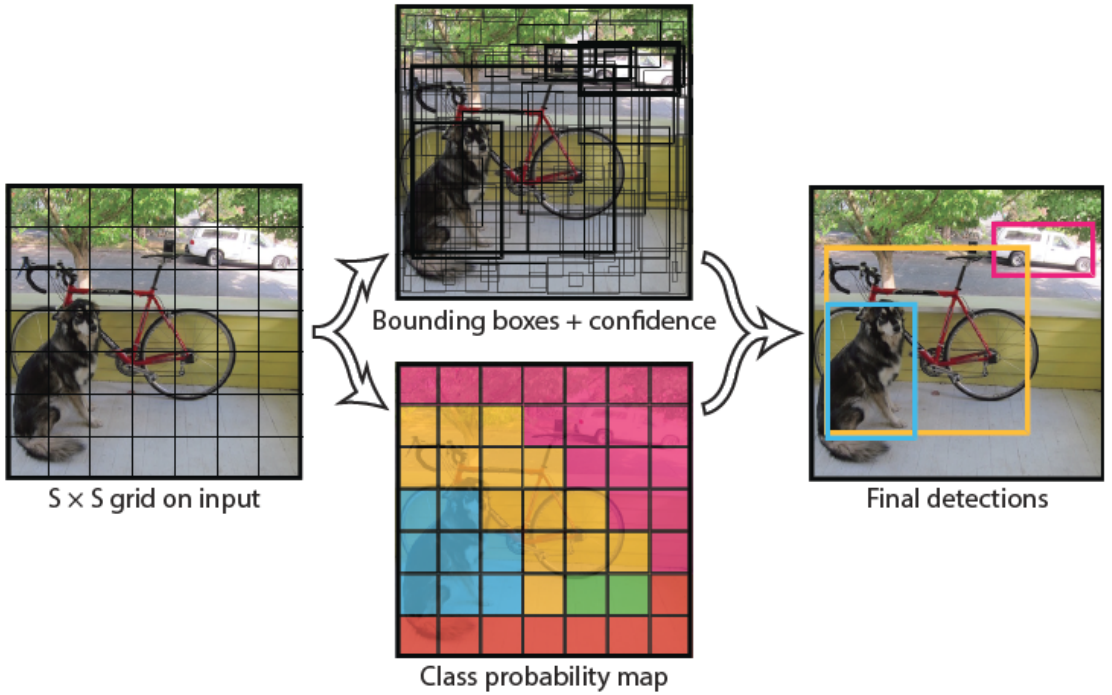


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

J. Redmon, et al. 2016. "You Only Look Once: Unified, Real-Time Object Detection". In 2016 CVPR.

Detection Process

- Divide input image into an $S \times S$ grid
- Each grid cell predicts B bounding boxes and confidence score for each box
  - No object means confidence score 0
- Each bounding box consists of 5 predictions: $(x, y, w, h, confidence)$
  - (x, y): Coordinates of center of the box
  - (w, h): Width and height of the box
- Each grid cell predicts C conditional class probabilities

Architecture

YOLOv1

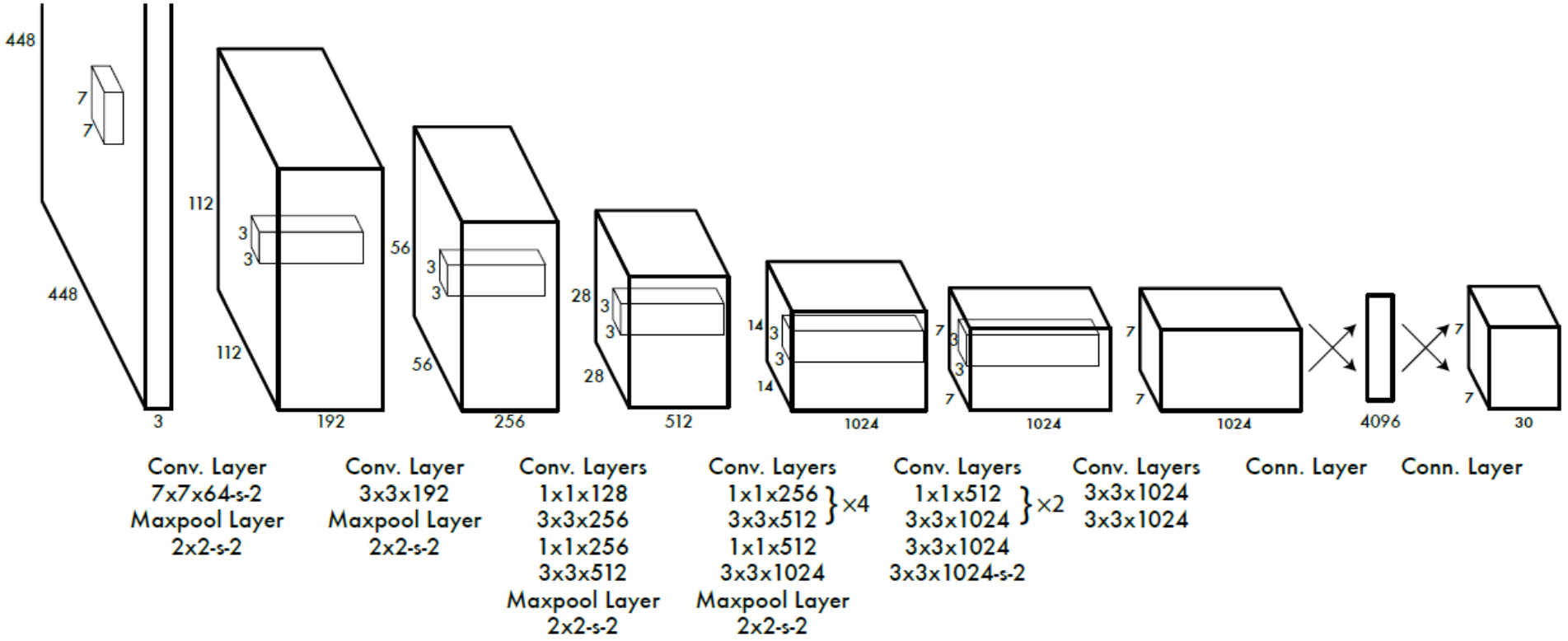


Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

J. Redmon, et al. 2016. "You Only Look Once: Unified, Real-Time Object Detection". In 2016 CVPR.

Quantitative Results

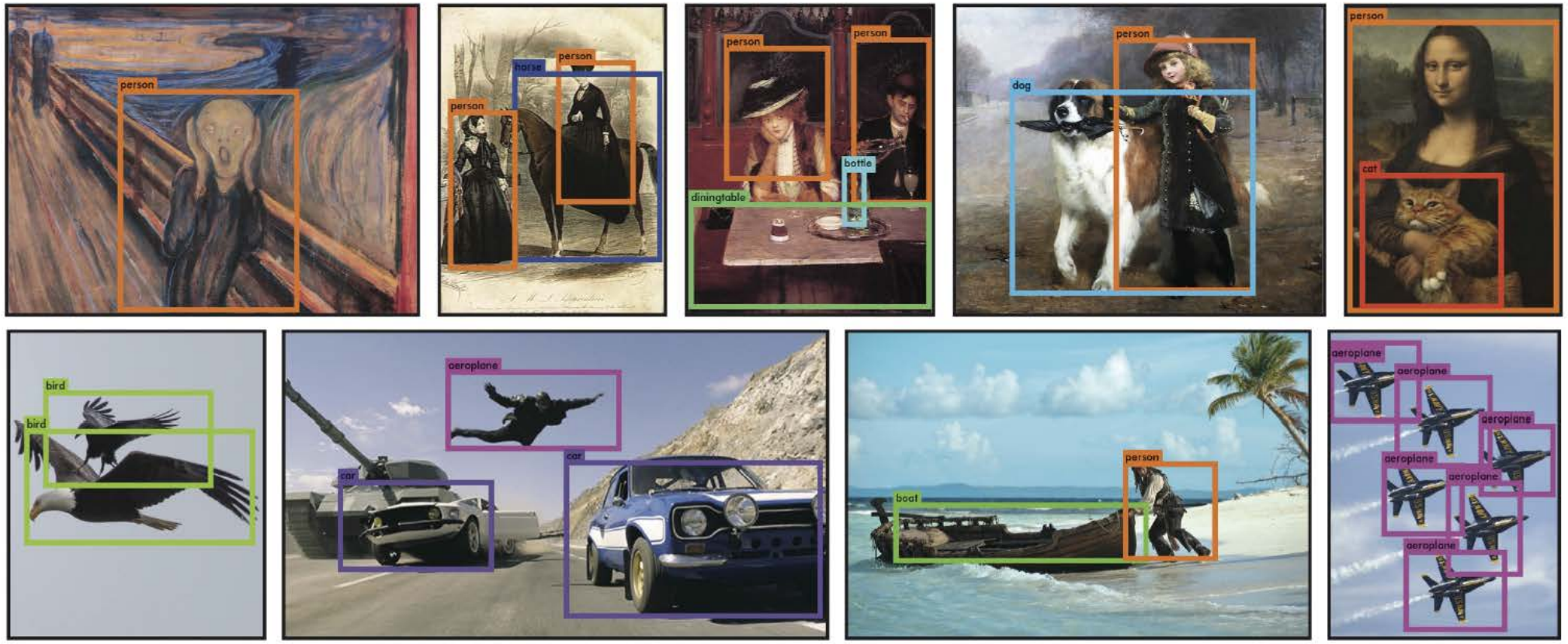


Figure 6: Qualitative Results. YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

J. Redmon, et al. 2016. "You Only Look Once: Unified, Real-Time Object Detection". In 2016 CVPR.

Limitations

- Small number of nearby objects that can be predicted
- Low performance detecting small objects
- Low performance detecting objects in new or unusual aspect ratios or configurations

Contributions

- Trained directly on full images
- Trained on a loss function that directly corresponds to detection performance
- Unified architecture that trains end-to-end jointly
- Object detection in real time
- Robust and generalized to new domains
- Simple to construct