Single Shot Object Detection in Aerial Images

Presented By:
Mohd Haroon Ansari
Pooja Gupta
Maj. Govind

Presented To:
Dr. Ambedkar Dukkipati

AGENDA

- Problem
- Recent Work
- State of the Art
- Overview of the Components
- Details of the Work
- Results
- References

Problem Statement

- **Geospatial object detection in Satellite Images**
- **Motivation:**
  - Intrusion detection in Defence
  - Automatic Identification System(AIS)
  - Traffic Analysis
  - Many more applications

Recent Work

- **R-CNN**
  - Use CNN for feature representation
  - Use SVM for classification
- **Faster R-CNN**
  - Use RPN for region proposal
  - Use R-CNN for object detection
- **YOLO**
  - Frame object detection as regression problem
  - Detect object using single neural network
- **SSD**
  - Predicts score of presence of object in box
  - Handles objects of various size

State of the Art

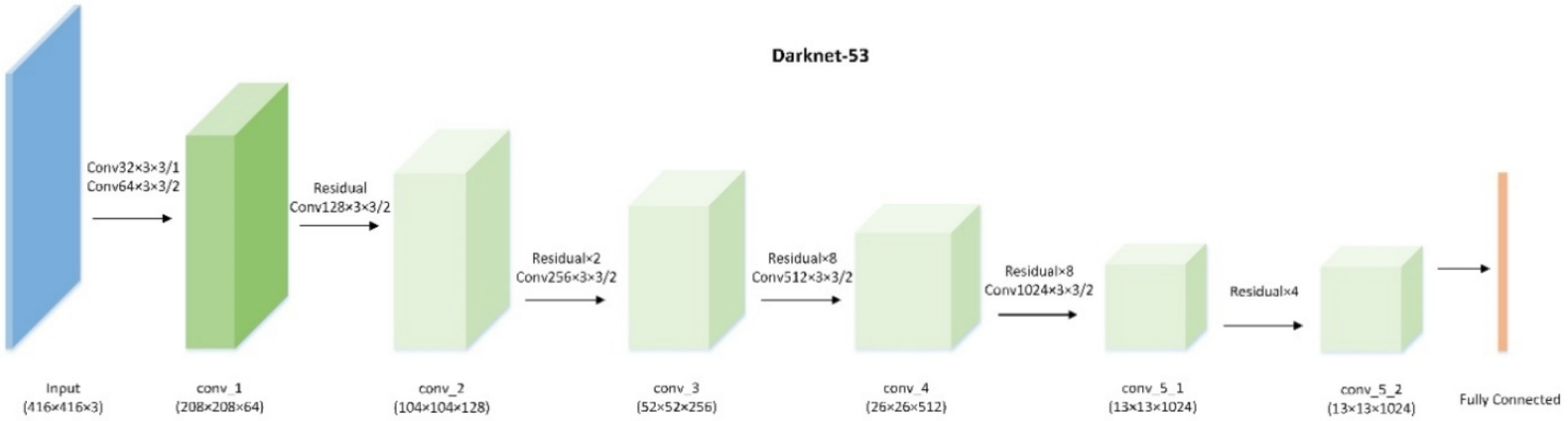
- **Single Shot Detection with Multiple Feature Fusion**
- **Use concepts from YOLO an SSD**
- **Faster as compared to YOLO and SSD**
- **Accuracy better than YOLO and SSD**

Overview of the Components

- **Base Feature Extractor**
- **Multi-Scale Feature Fusion Detector**
- **Multi Scale Feature Fusion Module**
- **Anchor Priors and Predictions**

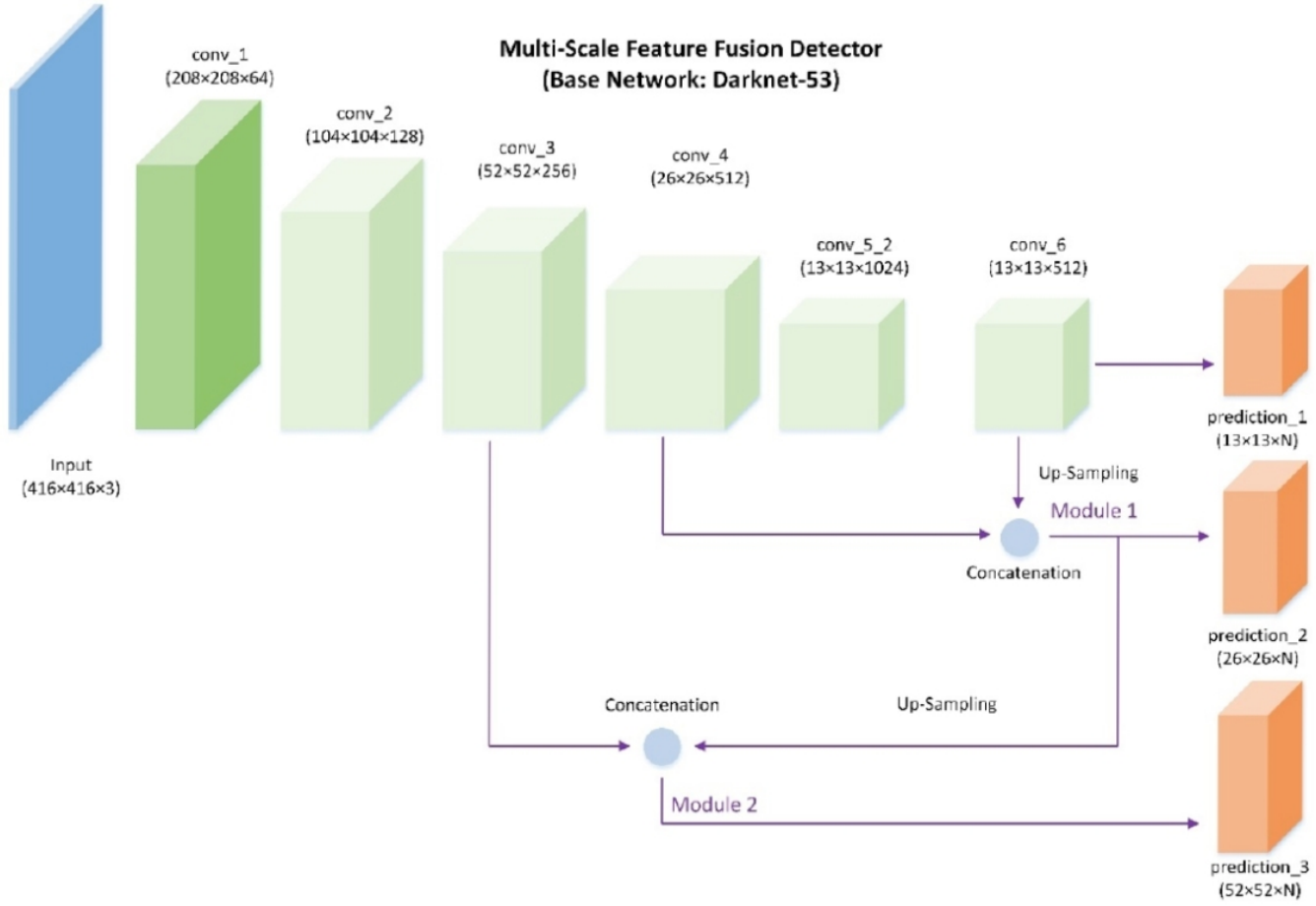
Base Feature Extractor

- **DarkNet-53**
  - 53 Convolutional Layer without pooling any layer
  - 23 residual blocks
  - Use Leaky ReLU as activation function

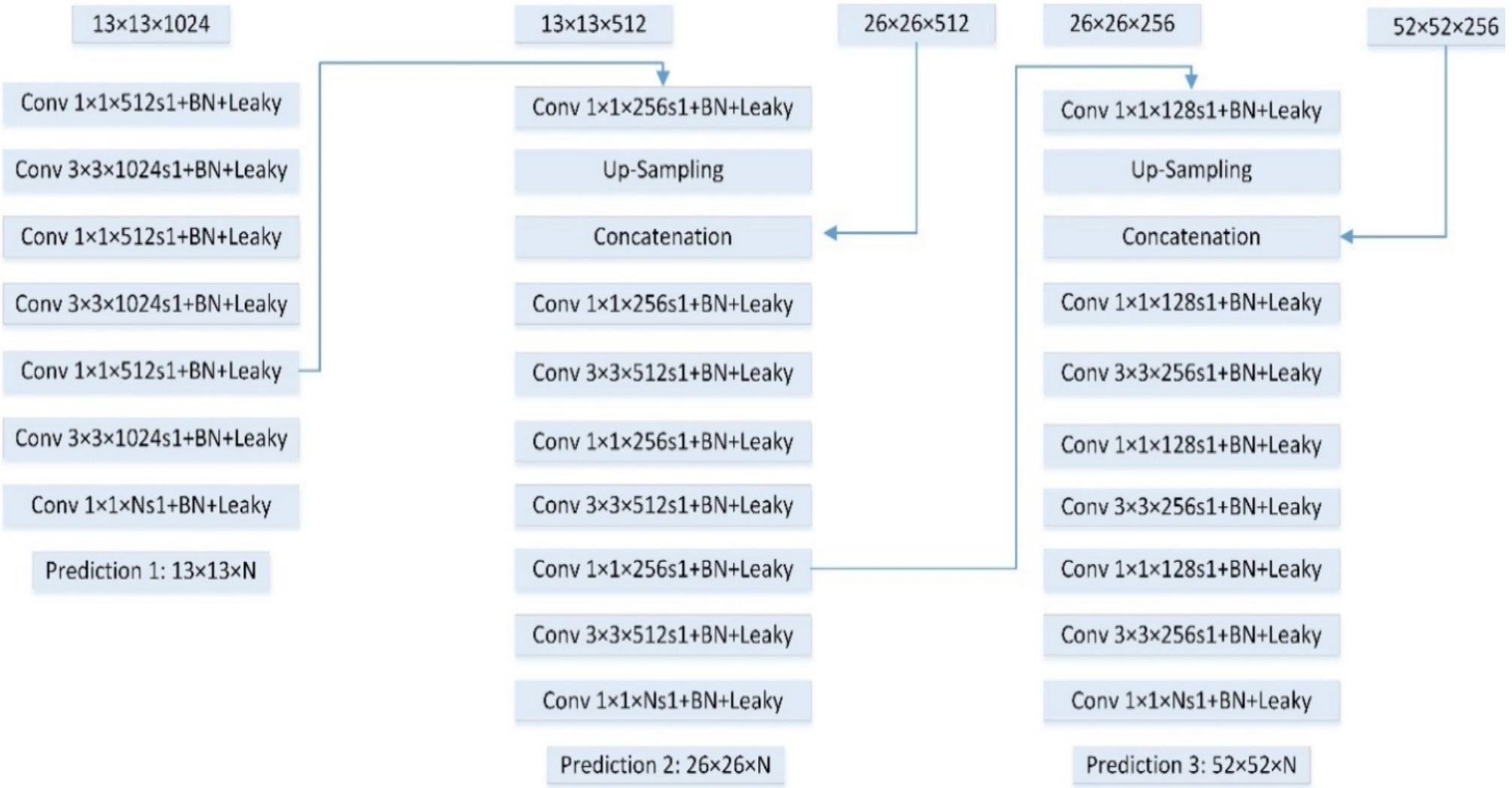


Multi Scale Feature Fusion Detector

- **Combines Features of Different Scale**
- **Predictions on different scales are made**

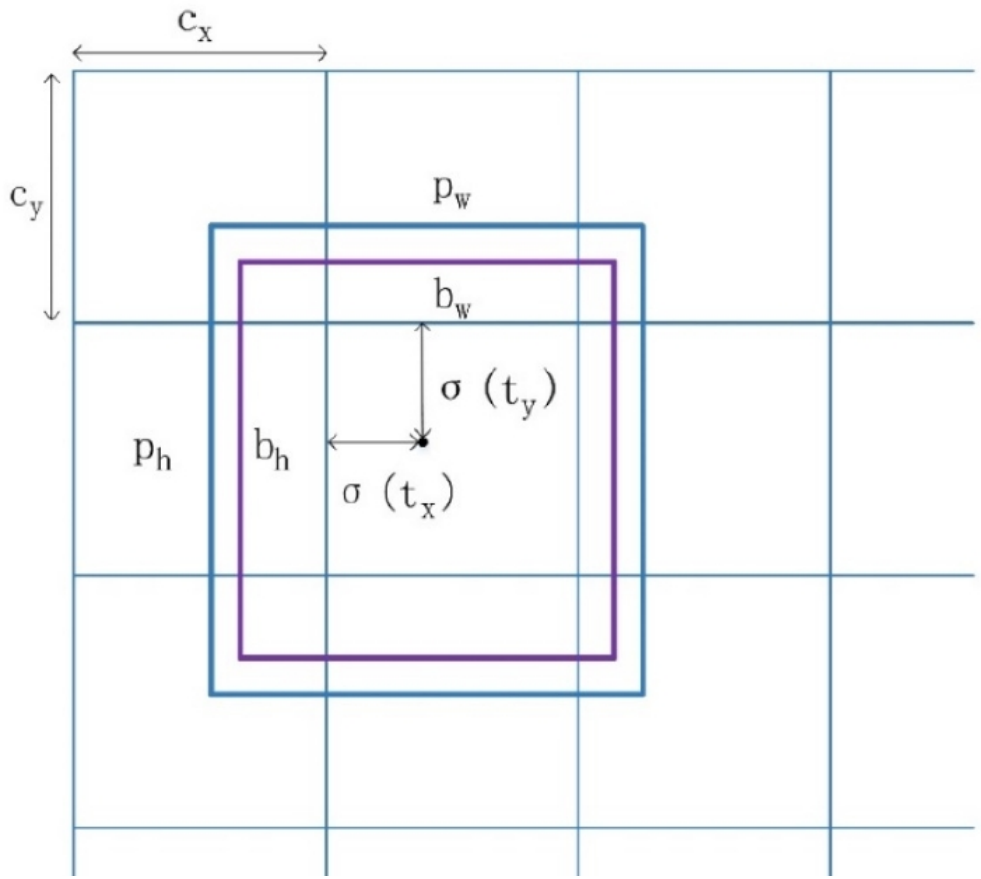


Multi Scale Feature Fusion Module



Anchor Prior and Predictions

- 3 Anchor priors are used for each of three predictions
- Anchor priors obtained using k-means clustering



$$b_x = \sigma\ (t_x) + c_x$$
$$b_y = \sigma\ (t_y) + c_y$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$

Results

- **Achieves high accuracy as compared to recently proposed models**

Method	Faster R-CNN	SSD	YOLO2	Proposed	Proposed (Soft NMS)
Pretrained backbone	ResNet50	VGG16	Darknet-19	Darknet-53	Darknet-53
Airport	**0.911**	0.788	0.598	0.839	0.847
Helicopter	0.876	0.893	0.917	0.946	**0.946**
Plane	0.673	0.819	0.813	0.897	**0.904**
Oiltank	0.645	0.898	0.909	0.920	**0.922**
Warship	0.759	0.755	0.695	0.793	**0.826**
Mean AP	0.773	0.831	0.786	0.879	**0.890**

References

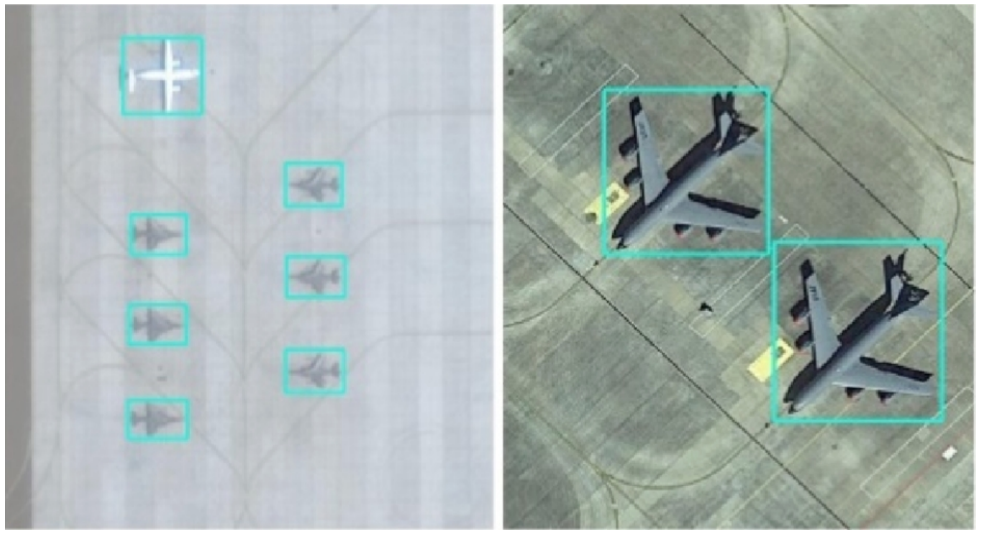
- Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv, 2013; arXiv:1311.2524.
- Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.
- Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. arXiv, 2015; arXiv:1506.02640.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector.
- Zhuang, Shuo, et al. "A Single Shot Framework with Multi-Scale Feature Fusion for Geospatial Object Detection." Remote Sensing 11.5 (2019): 594.

Thank You

Images with their the corresponding annotated bounding boxes

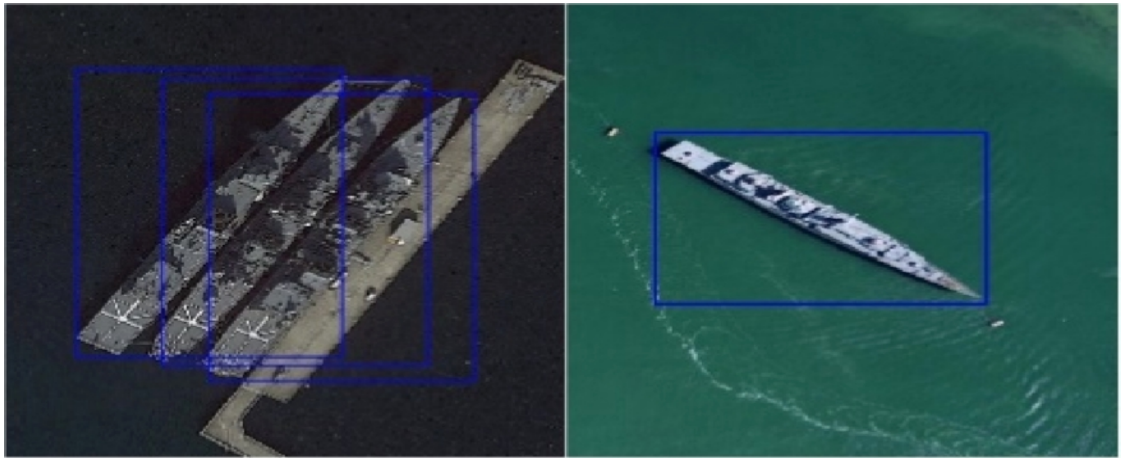
Airport

Plane



Oiltank

Warship

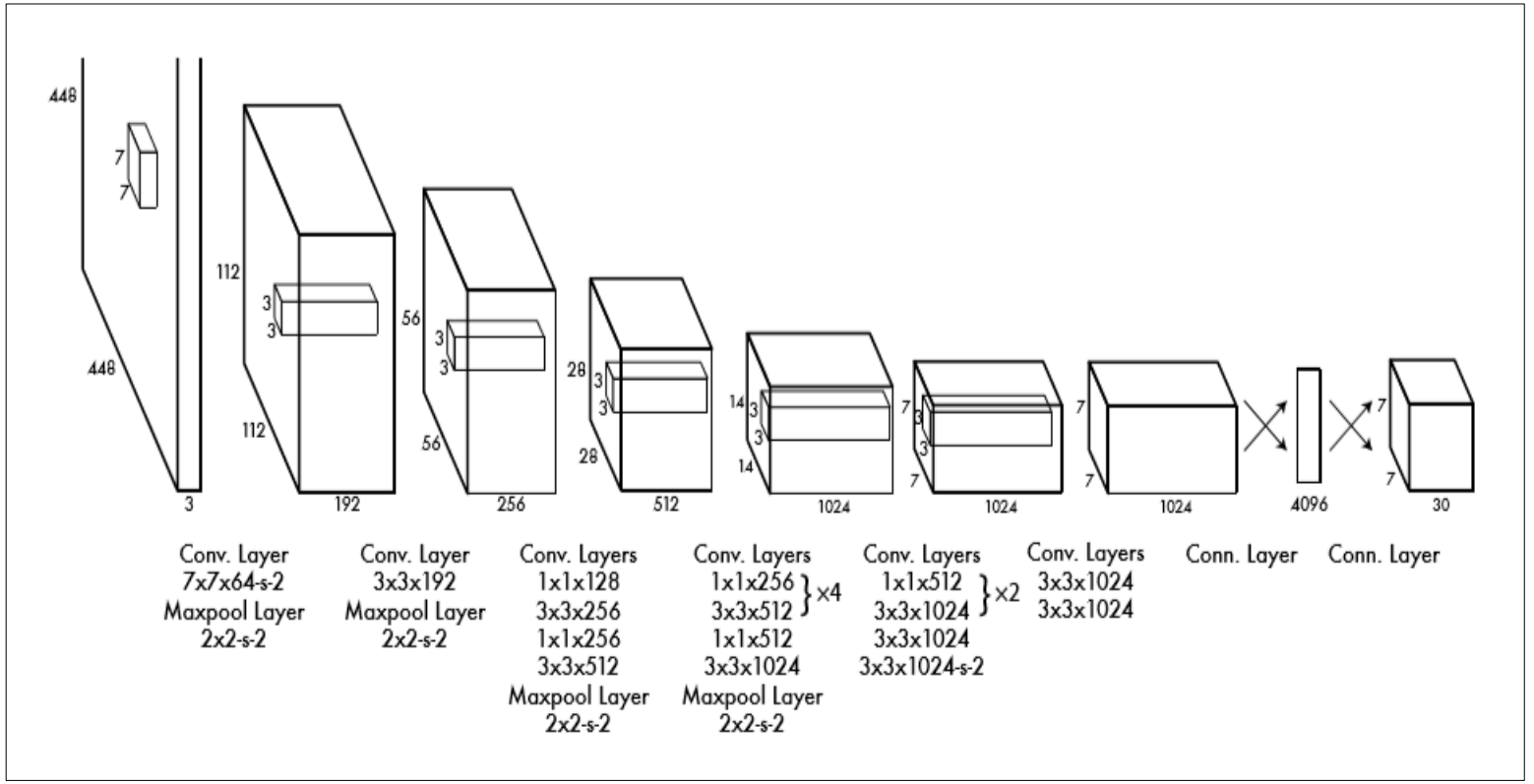


Helicopter

Remote-Sensing Dataset for Geospatial Object Detection (RSD-GOD)

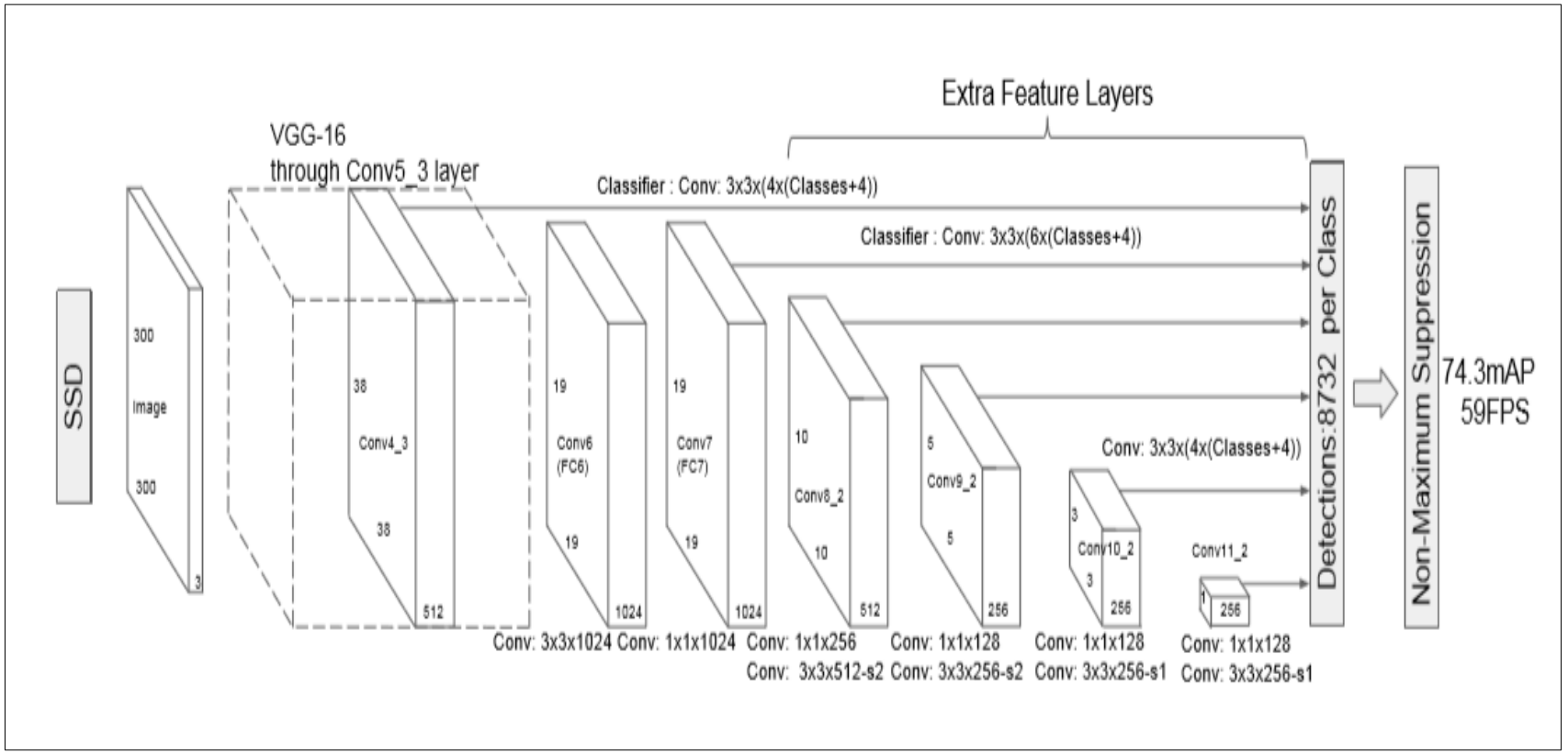
- RSD-GOD is a large scale dataset for , that consists of 5 different categories with **18,187 annotated images and 40,990 instances**.
- Considering the applications in **military field**, five categories are selected to be annotated, including **plane, helicopter, oil tank, airport and warship.**
- RSD satisfies following three properties :
  - Rich background information
  - Multiple resolutions and viewpoints
  - Dense objects.

You Only Look Once(YOLO)



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.

Single Shot MultiBox Detect(SSD)



Architecture of SSD compared to YOLO : It adds several feature layers to the end of a base network, which predict the offsets to default boxes of different scales and aspect ratios and their associated confidences. SSD with a 300×300 input size significantly outperforms its 448×448 YOLO counterpart in accuracy on VOC2007 test while also improving the speed.

Loss Function

$$L_{overall} = L_{loc} + L_{conf} + L_{cla}$$

$$L_{loc} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (w_i - \hat{w}_i)^2 + \left(h_i - \hat{h}_i \right)^2 \right]$$

$$L_{conf} = \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} P^{obj} (c_i - \hat{c}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \left(1 - P^{obj} \right) (c_i - \hat{c}_i)^2$$

$$L_{cla} = -\lambda_{cla} \sum_{i=0}^{S^2} P^{obj} \log(\hat{p}_i)$$