CVPR 2018 WAD Video Segmentation Challenge

Azat Akhtyamov
azakhtyamov@yandex.ru

Problem Description

- Instance Segmentation
- 7 classes:
  - Car
  - Motorcycle
  - Bicycle
  - Pedestrian
  - Truck
  - Bus
  - Tricycle (???)
- Groups and riders are ignored
- 2 submission per day
- 100 GB train, 4.2 GB test
- 3384x2710 each image ---> 3384x1536
- 3 videos, ~7 fps
- Evaluation Metric: mean average precision (mAP)

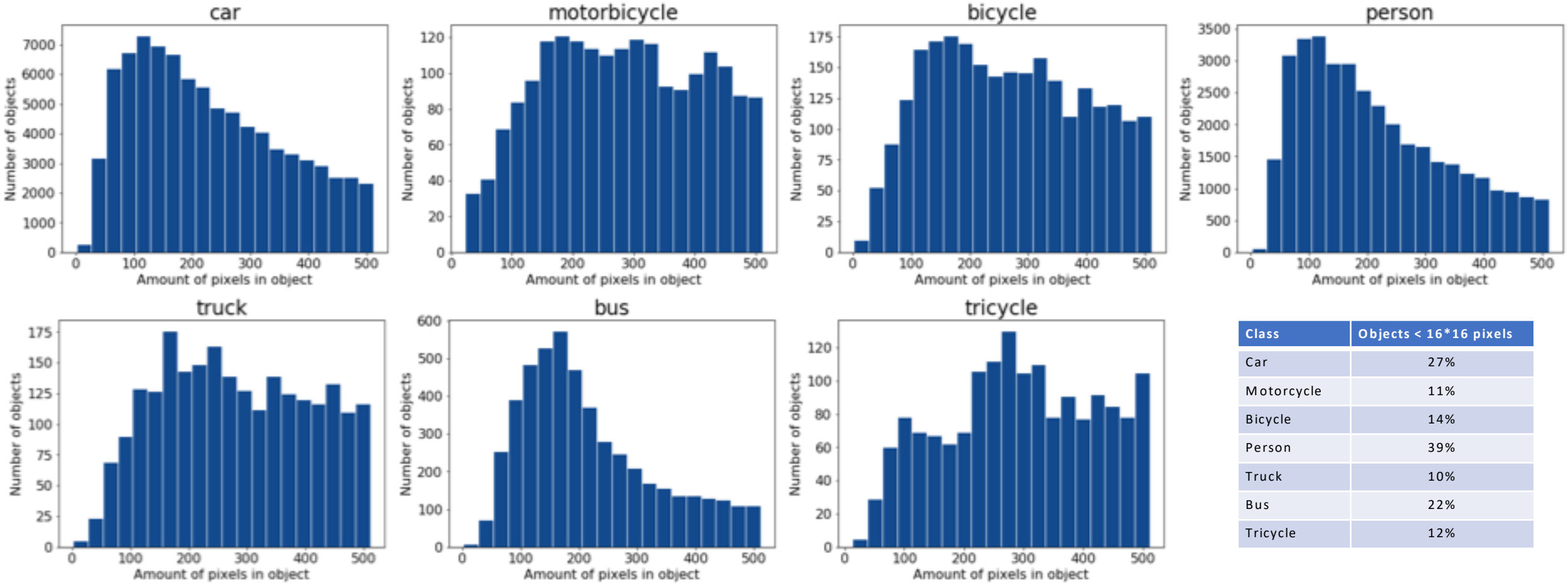
Evaluation Metric: mAP

A – prediction mask, B – ground truth mask

- $IoU(A, B) = \frac{A \cap B}{A \cup B}$
- TP_τ – if $IoU(A_i, B) \geq \tau$ and $i == \underset{i}{\text{argmax}}\, Conf(A_i, B)$
- FP_τ – if $IoU(A_i, B) < \tau$ or $i \neq \underset{i}{\text{argmax}}\, Conf(A_i, B)$
- FN_τ – if $IoU(A_i, B_j) < \tau, \forall i$
- $AP = \frac{1}{10} \sum_\tau \frac{TP_\tau}{TP_\tau + FP_\tau + FN_\tau}, \tau \in [0.5, 0.55, \ldots 0.9, 0.95]$
- mAP – mean AP over classes and clips

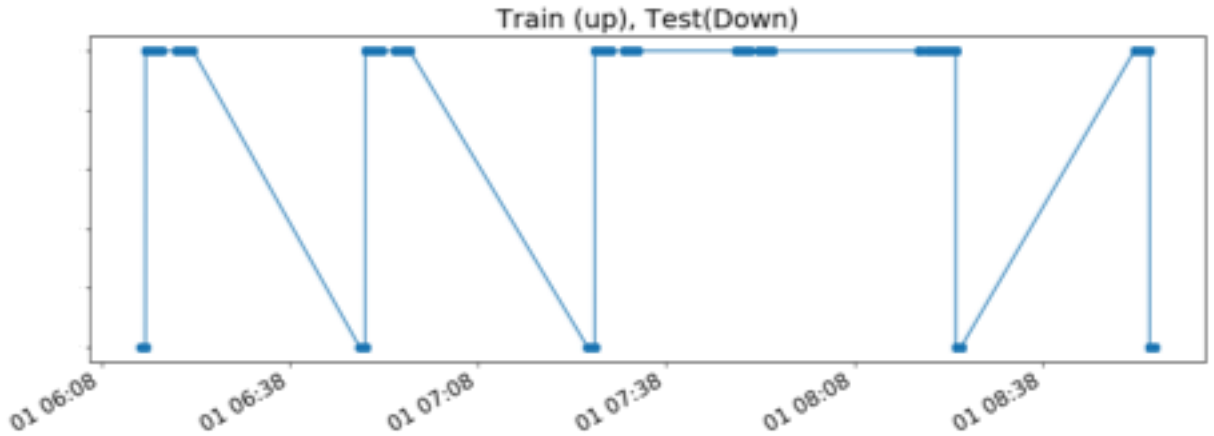
Original: https://medium.com/@timothycarlen/understanding-the-map-evaluation-metric-for-object-detection-a07fe6962cf3

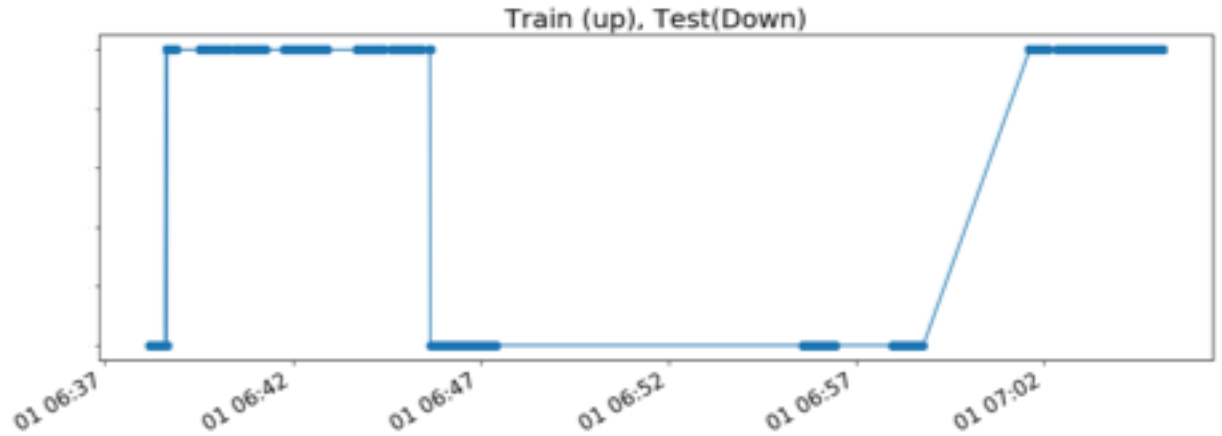
Dataset: Masks' Sizes

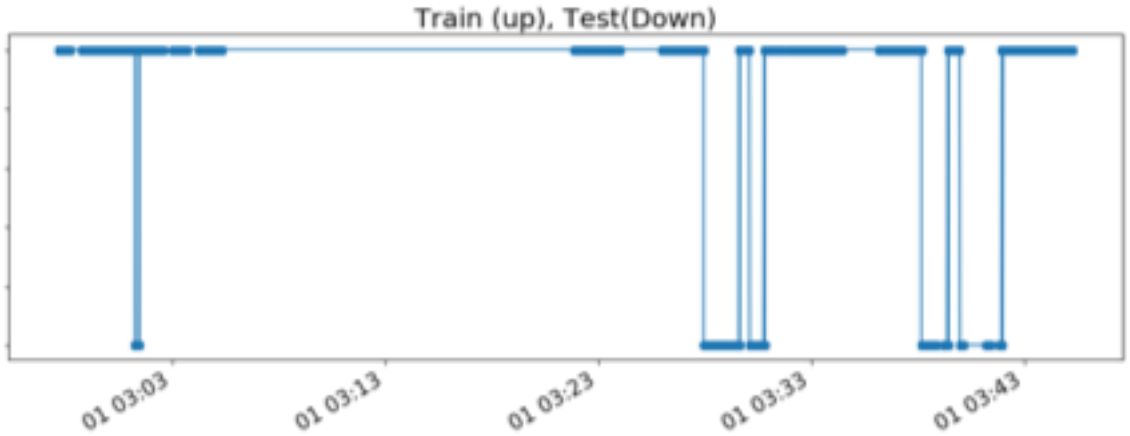


Class	Objects < 16*16 pixels
Car	27%
Motorcycle	11%
Bicycle	14%
Person	39%
Truck	10%
Bus	22%
Tricycle	12%

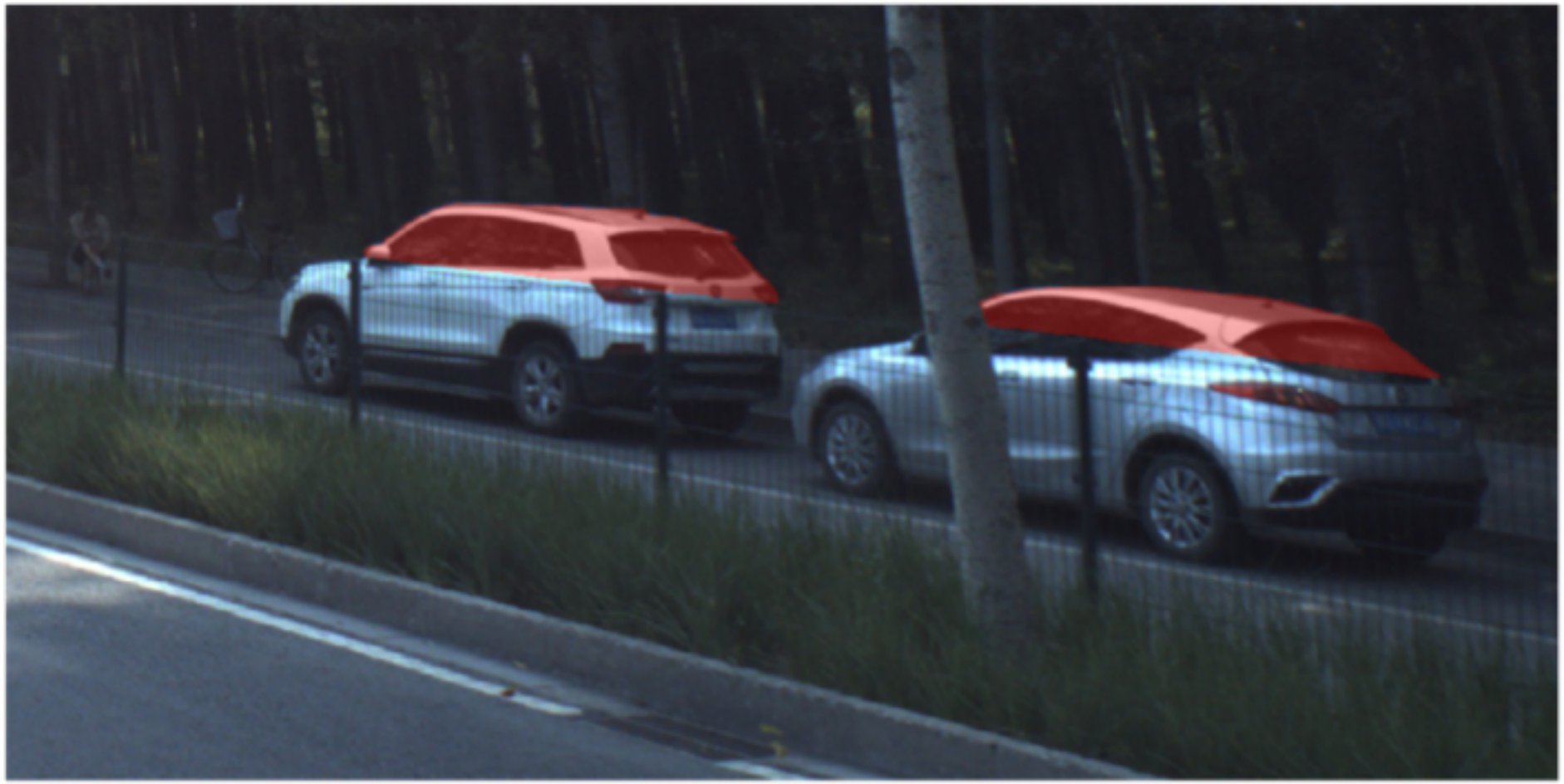
Dataset: Train/Test Timeline





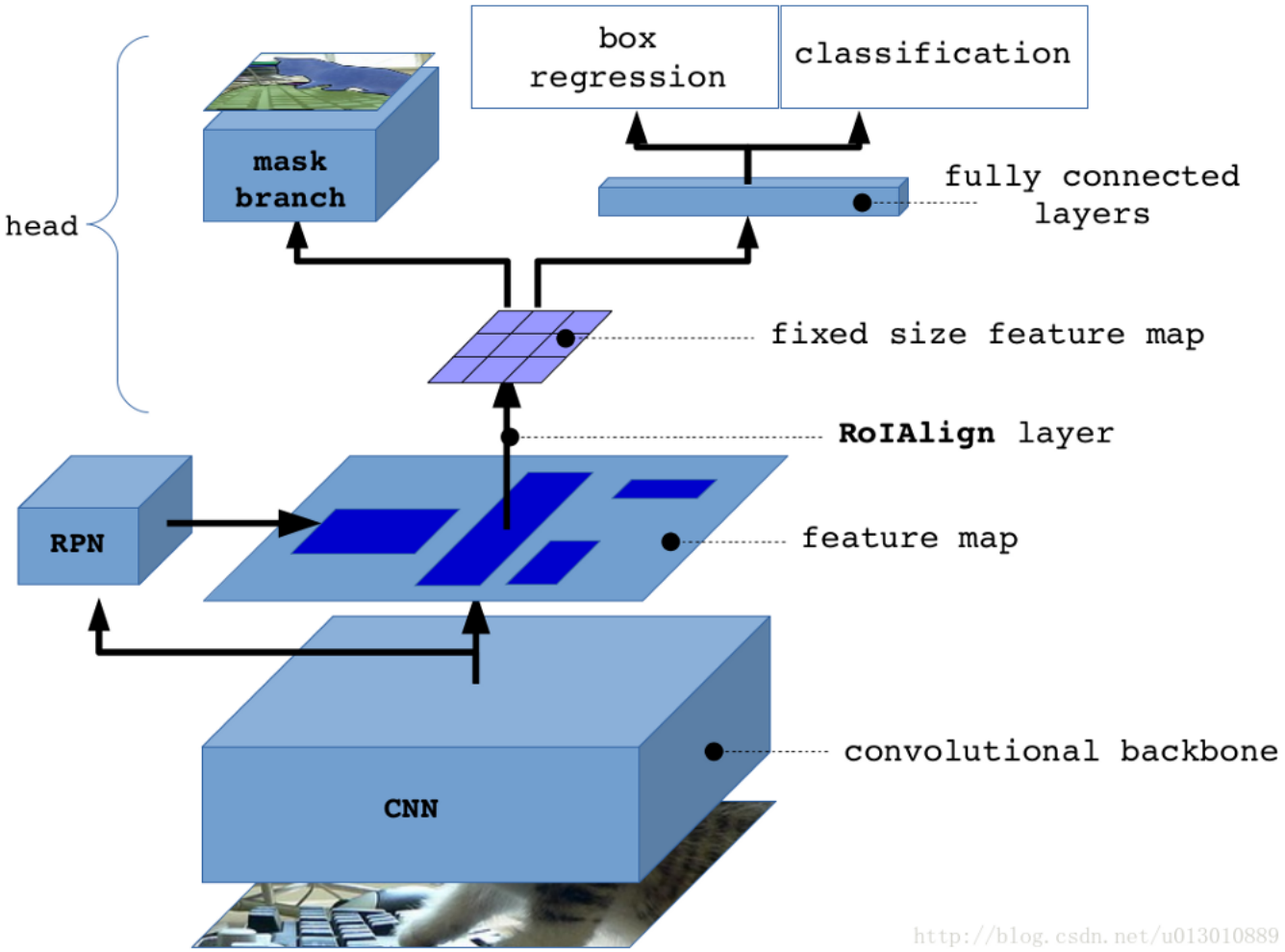


Dataset: Quality



Dataset: Quality

Mask R-CNN



$$L = L_{cls} + L_{box} + L_{mask}$$

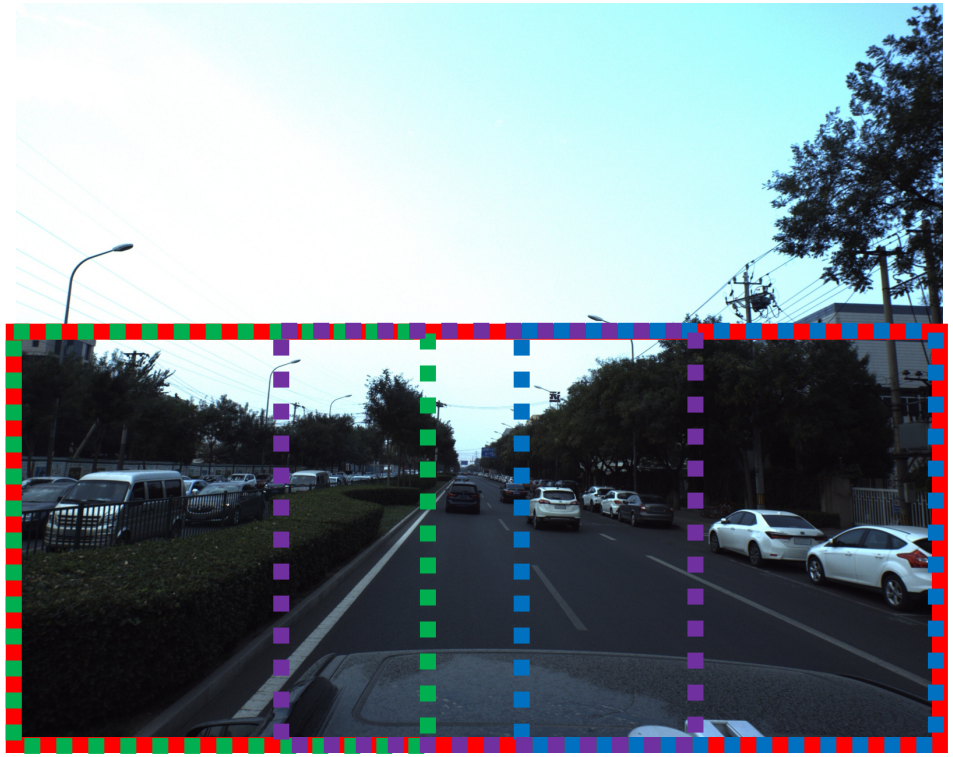
Original to: https://www.slideshare.net/IldooKim/deep-object-detectors-1-20166

Pipeline: Training and Prediction

- Training with random crops 1536x1536, scaling from 50% to 200%
- Augmentations (brightness and contrast)
- Test set overfitting
- 4 predictions: all, left, middle, right
  - Add everything from middle
  - For [left, right]:

    If intersects with added – add from all

    Else add



Pipeline: Deleting Masks

Predict if mask should be deleted, given context and other masks:

- Resnet101
- Input: [grayscale image, current mask, other masks]
- Delete if $p > 0.9$ (3% of predictions))

Pipeline: Handling Intersections

- Heuristics:
  1) Different classes:
     Pedestrian > Car
     Car > Bus
     …
  2) Same class: assign intersection to the object with higher confidence

Total inference time: 12h
Val mAP: ~0.3

Results

Big and square – it's a bus!