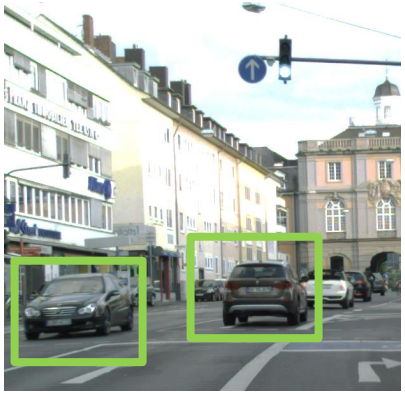
Photogrammetry & Robotics Lab

Machine Learning for Robotics and Computer Vision Tutorial

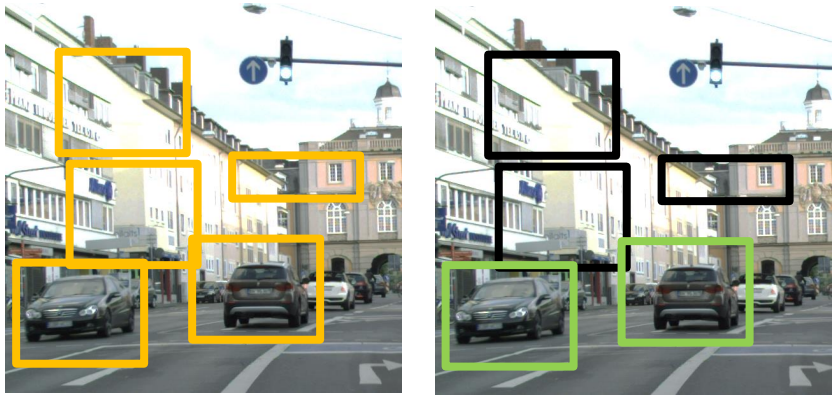
Detection with CNNs

Jens Behley

Single vs. Two-Stage Approaches



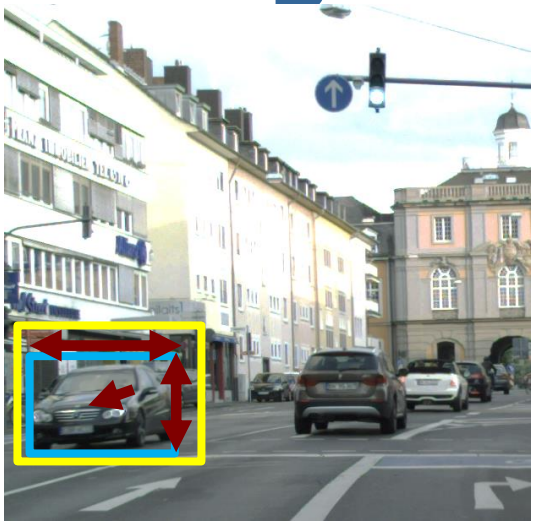
Single-stage

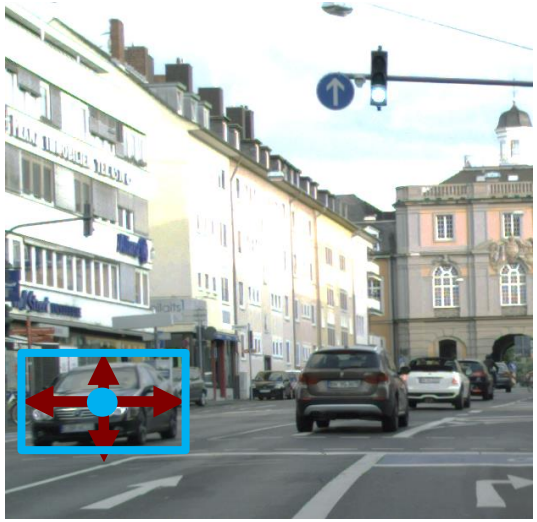


Two-stage

- Two paradigms for Object Detection:
  1. **Single-stage approaches**: Directly produces bounding boxes in single forward pass
  2. **Two-stage approaches:** First generates class-agnostic proposals and classifies only top N-proposals

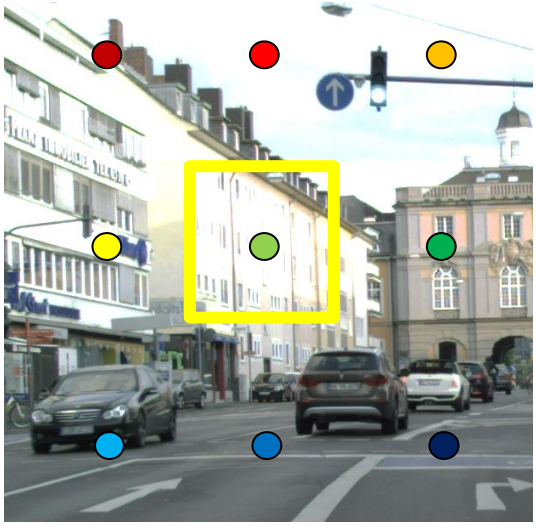
How to get a CNN to output bounding boxes?

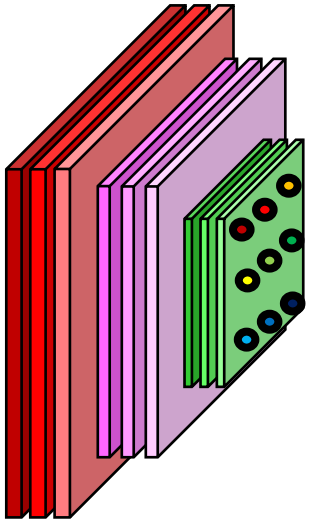




- **Anchor-based approach**: provide templates that need to be classified and "modified"
  - Examples: R-CNN, Fast R-CNN, Faster R-CNN, YOLO, EfficientDet, FPN, RetinaNet
- **Anchor-free approach**: produce corners or centers (key point) that produces the desired bounding box
  - Examples: CornerNet, CenterNet

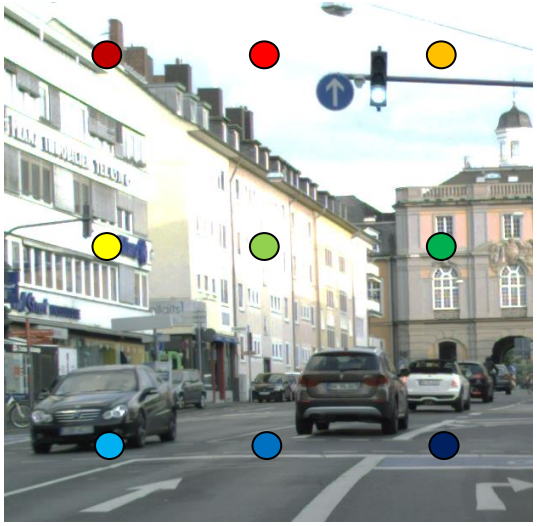
Anchor-Based Approach

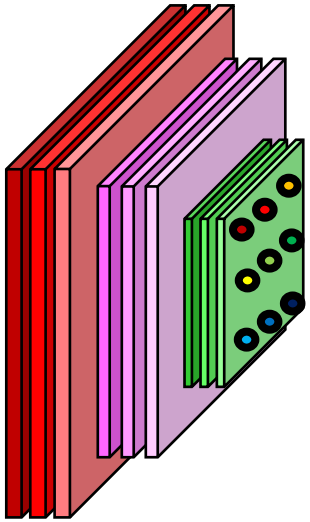




- Each location in feature map corresponds to spatial position in the image

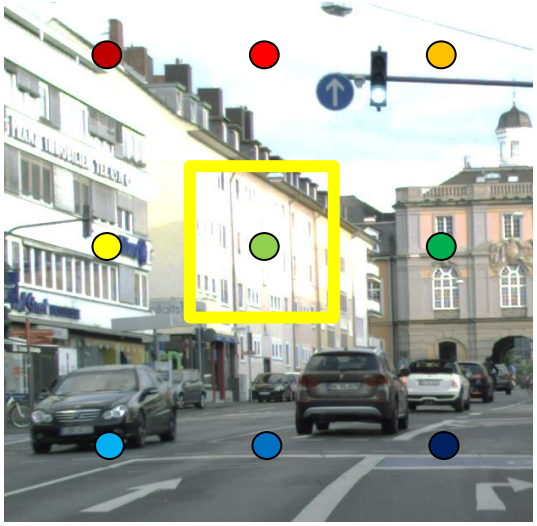
Anchor-Based Approach

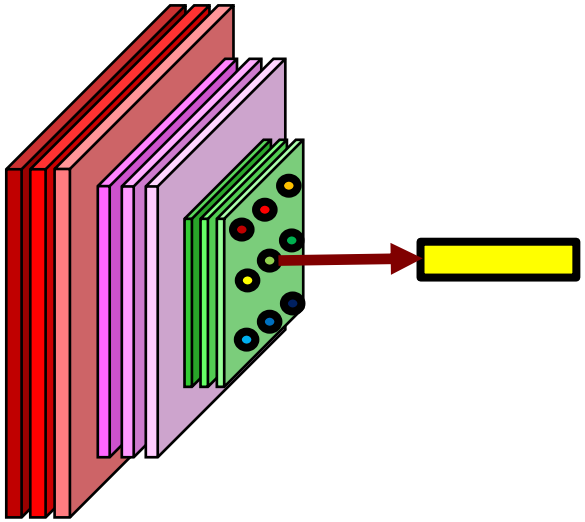




- At each location are anchors located
- Different aspect ratios, different sizes

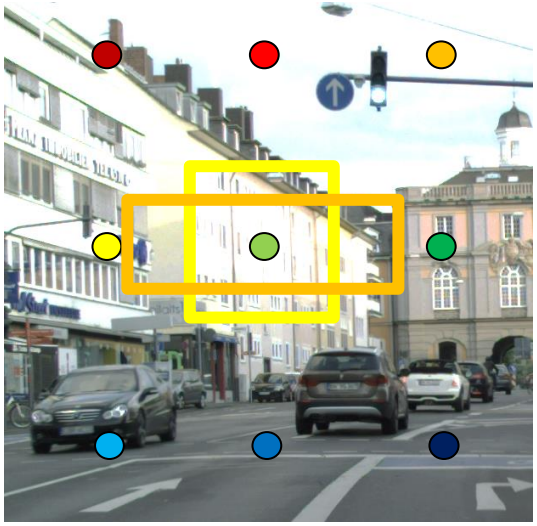
Anchor-Based Approach

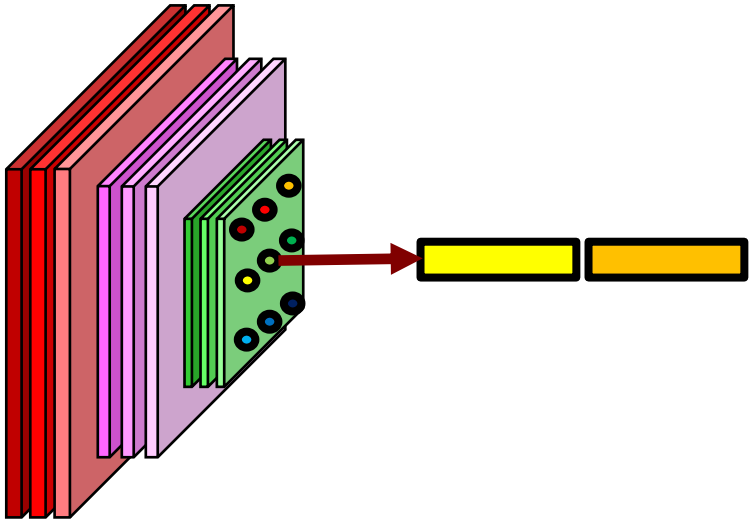




- For each anchor, we produce class + bbox offsets

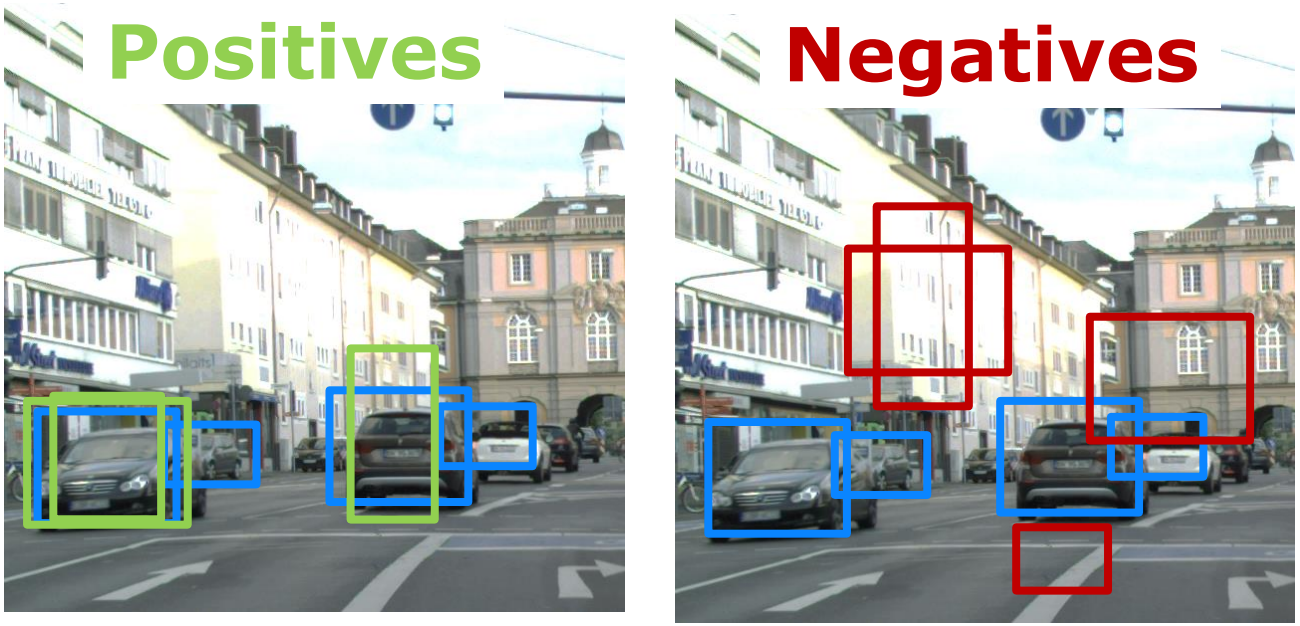
Anchor-Based Approach





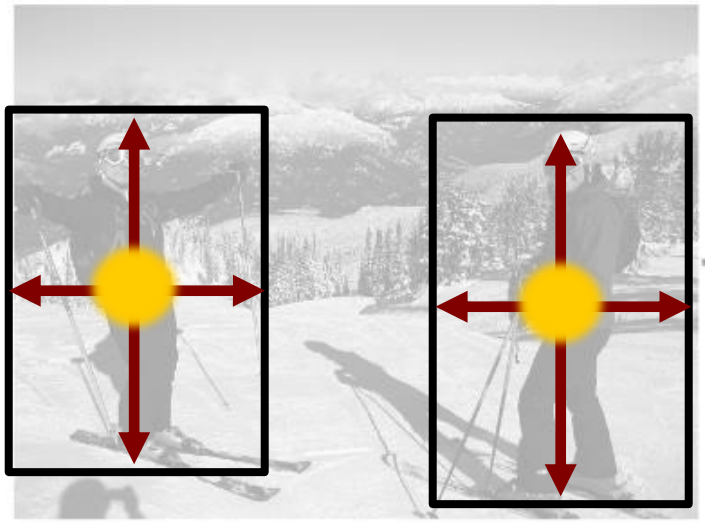
- For each anchor, we produce class + bbox offsets

Anchor Assignment



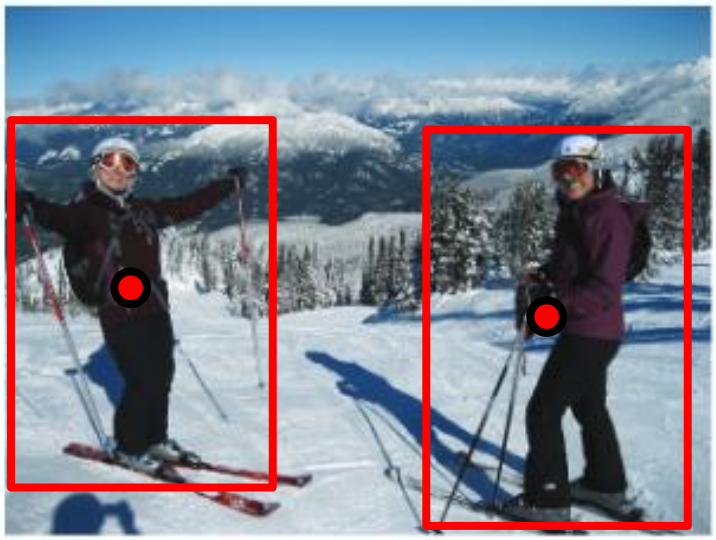
- IoU-based assignment to determine positive vs. negative examples
  - Positive: highest or IoU > 0.7 with ground truth box
  - Negative: IoU < 0.3 for all ground truth boxes
- Usually far more negatives than positive boxes

Anchor-free Approaches



- Produce for location in the image how likely is that there is a bounding box
- In case of CenterNet: Likelihood that for specific class at that location

Center Heatmap



Y

$$Y_{xyc} = \exp\left(-\frac{(x-c_x)^2+(y-c_y)^2}{2\sigma_p^2}\right)$$

$$Y \in \mathbb{R}^{H \times W \times K}$$

- Target for centerness heatmap is Gaussian at center location (variance is object size dependent)
- In case of overlap: maximum of values
- At center it is 1 and falls of with distance to the center

Center Loss

Focal loss

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}) & \text{if } Y_{xyc} = 1 \\ (1 - Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} & \text{otherwise} \\ \log(1 - \hat{Y}_{xyc}) & \end{cases}$$

- For a prediction $\hat{Y}$ the loss is now computed per-pixel-wise in respect to ground truth map Y
- For exact center location, we want prediction to be one
- For non-center locations, we want to push it to zero

Size estimation

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{p_k} - s_k \right|$$

- Here we want the size, e.g., width and height, predicted at the center location to be as close to real size
- Smoothed L1 loss (or L1 loss) used to compute the loss here (functional.smooth_l1_loss)

Complete Loss

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{off} L_{off}.$$

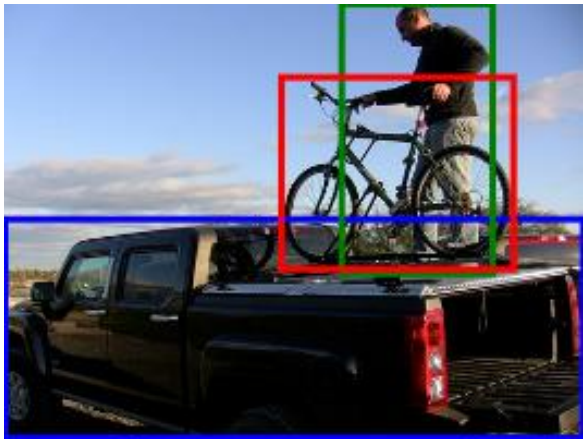
- Weighted sum of center loss, size loss, and offset loss.
- CenterNet uses 0.1 for size and 1 for offsets.

Extracting Bounding Boxes

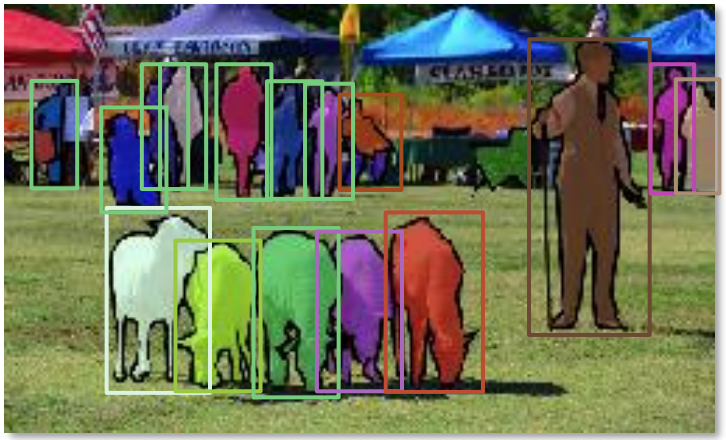
- Simple algorithm for getting bounding boxes
  - Find 100 peaks (maximum in 8x8 neighborhood → 3x3 maximum pooling) for each category
- Centerness from heatmap is detection score of the bounding

Object Detection Datasets

Pascal VOC



ImageNet



MS COCO

LVIS

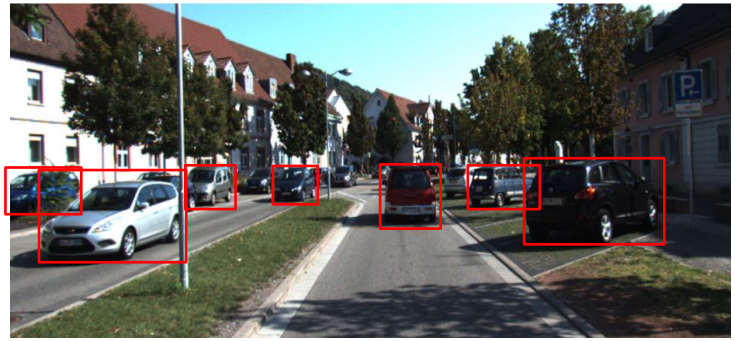
Dataset (Overview)

Name	Year	#Categories	#Images	Data
Pascal VOC	2012	20	12k	B
ImageNet	2014	200	477k	B
MS COCO	2014	80	123k	B, S
LVIS	2019	1000	164k	B, S
Objects365	2019	365	638k	B
Open Images	2020	600	1.9M	B

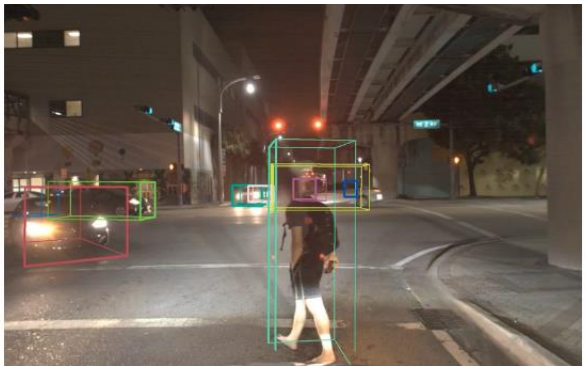
Bounding Box (B), Segmentation Masks (S)

- Images gather from Image Databases
- Mostly handheld cameras, smart phones

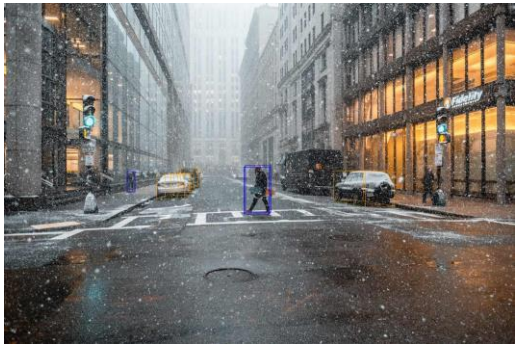
Automotive Datasets



KITTI



Argoverse



NuScenes

Waymo Open Dataset

Automotive Dataset (Overview)

Name	Year	#Categories	#Images	Data
KITTI	2012	8	15k	B
BDD100K	2017	10	100k	B
ApolloScape	2018	8-35	144k	B
KAIST	2018	3	9k	B
Argoverse	2019	15	22k	B
Lyft L5	2019	9	46k	B
A2D2	2019	14	12k	B
nuScenes	2019	23	40k	B,S
Waymo Open	2019	4	200k	B

Bounding Box (B), Segmentation Masks (S)

See you next week!