Object Detection

Philipp Krähenbühl, UT Austin

Image classification

Task

- Assign a single label to image

car

apple

Christmas tree

broccoli

Image Credit: https://en.wikipedia.org/wiki/File:Tiverton_Gazette_Newsroom.JPG

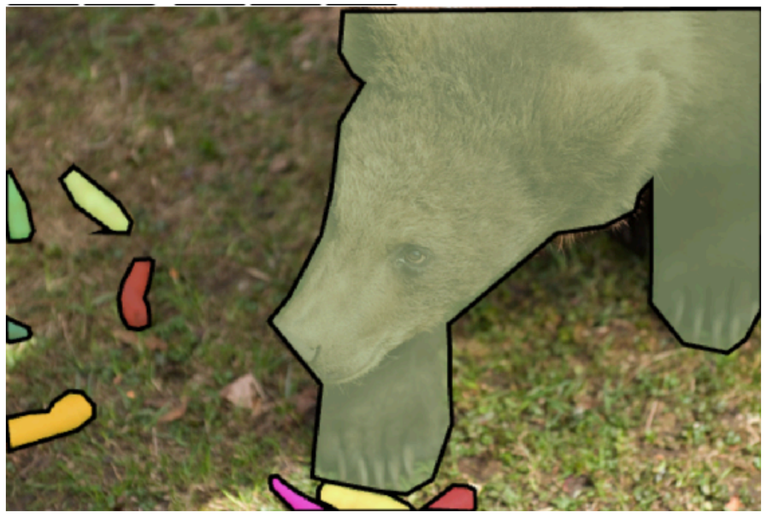
Credit: Frans de Waal https://www.youtube.com/watch?v=meiU6TxysCg

How do we reason about objects?

Object Detection - Datasets

MS COCO

- 120k training images
- 80 categories
- Box + segmentation annotations
  - For almost all objects



Microsoft coco: Common objects in context, Lin et al., 2014

Object detector

RCNN

Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al., 2014

Object detector

RCNN

- For any potential box
  - Heuristic: Object or not
  - Crop image
    - Classify
- Very slow

Object detector

FasterRCNN: 2 stage detectors

Faster R-CNN: Towards real-time object detection with region proposal networks, Ren et al., 2017

Object detector

FasterRCNN: 2 stage detectors

- Encode image using CNN
- For every pixel / patch enumerate n boxes
  - Predict “objectness”
  - Crop feature map
    - Classify
- Fast
- not end-to-end

Object detector

YOLO

You only look once: Unified, real-time object detection, Redmon et al., 2016

Object detector

YOLO: 1 stage detectors

- Encode image using CNN
- For every pixel / patch enumerate n boxes
  - Predict class or background
- Faster
- Almost end-to-end

Why do we use boxes?

- Reduction to image classification
  - Image classifications works very well
- Easy to annotate
- Decent distance measure
  - Overlap

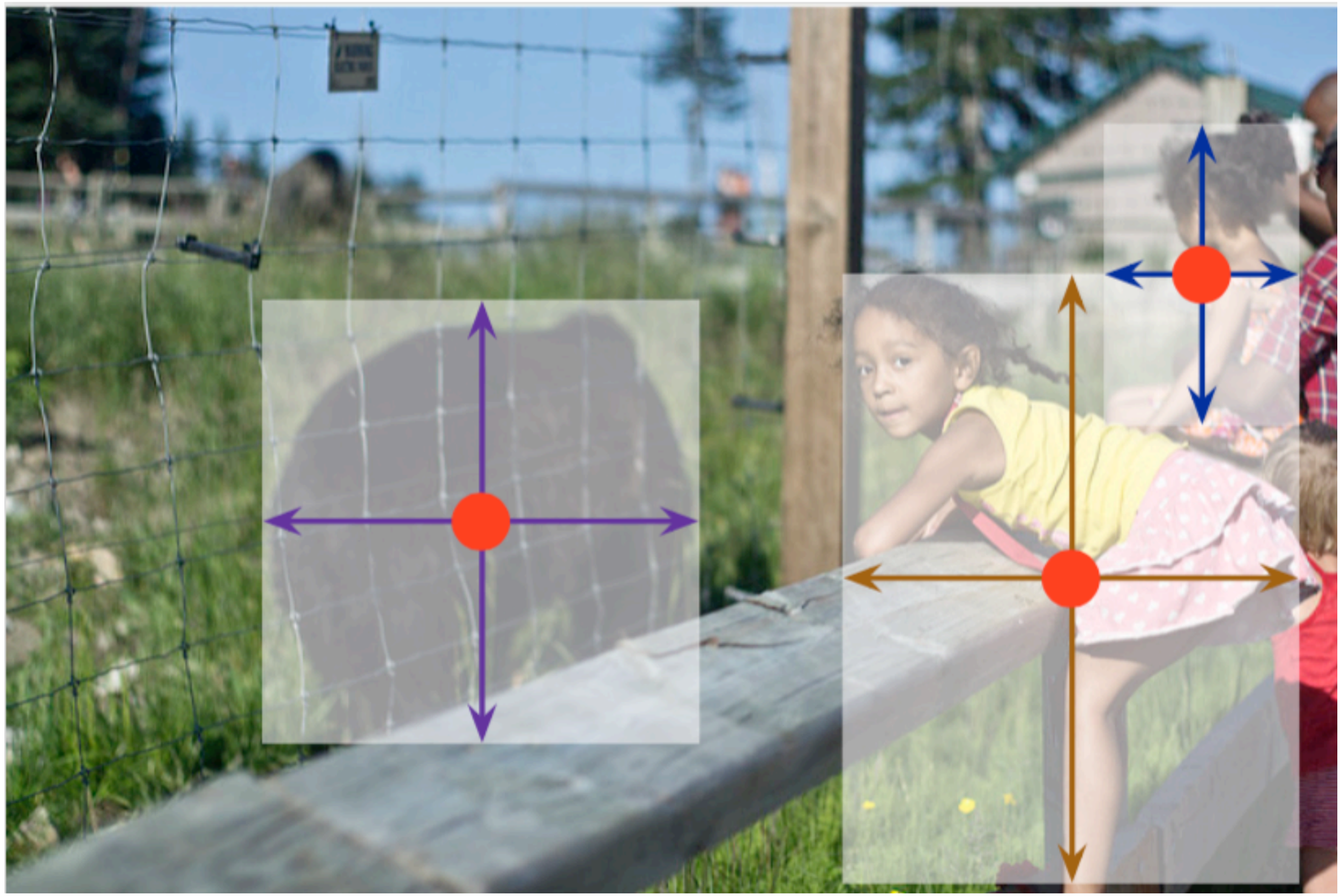
Object detector

Reasons not to use boxes

- Too many boxes $O(W^2H^2)$
  - Anchors and assignment (training)
  - Non-maxima suppression (testing)
  - Easy to miss oddly shaped objects

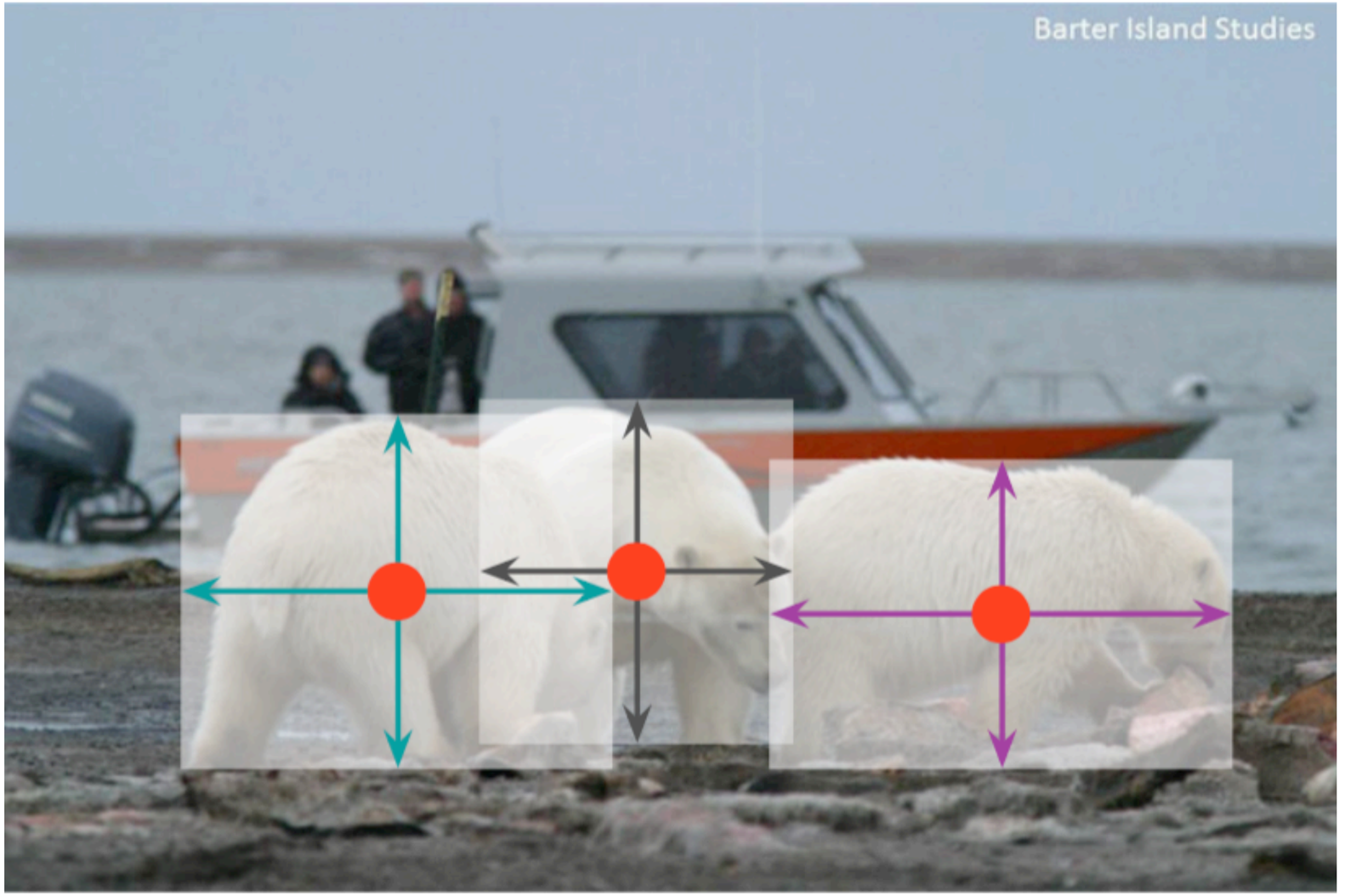
Simpler object detection

Objects as points



Objects as Points, Zhou etal., 2019

Objects as points



Objects as Points

CNN

Object Detector

Objects as Points

- Detect center points
  - Predict class, width and height
- Fast and accurate
- Almost end-to-end

Object detector

DETR

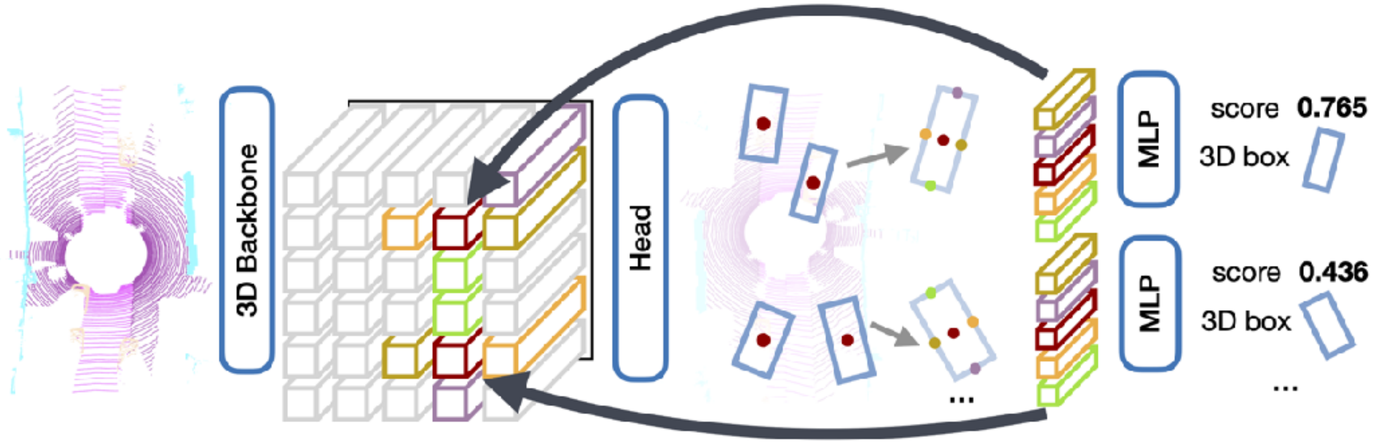
Object detector

DETR

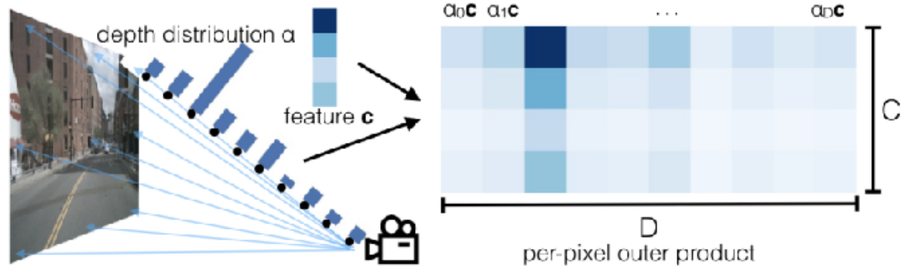
- Encode image using CNN or ViT
- Define Object Queries
  - Transform them into detections
    - Cross attend to image
- First end-to-end detector

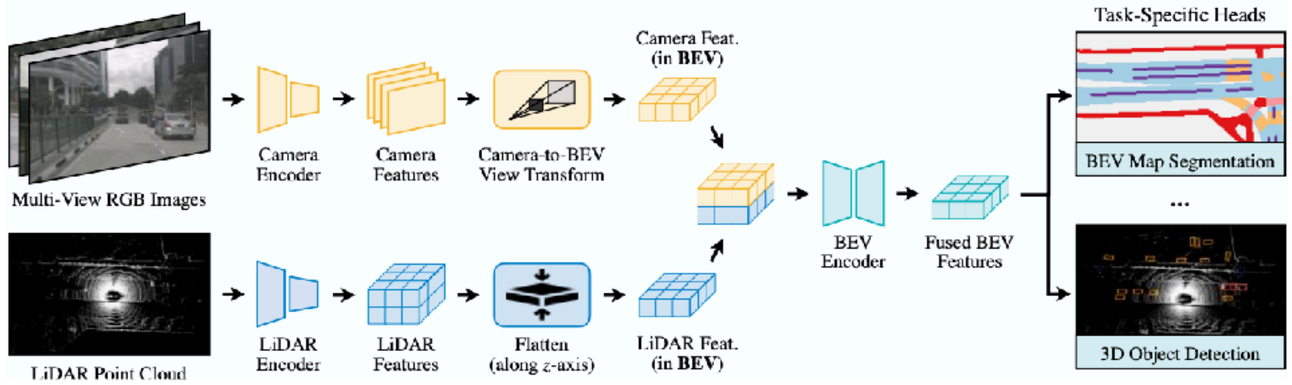
End-to-end object detection with transformers, Carion et al., 2020

3D Detection



- Option 1: Top-Down 2D Detection
  - Often relies on LiDAR sensor
- Option 2: Depth prediction
  - Project 2D detections to 3D using depth prediction





Center-based 3D Object Detection and Tracking, Yin etal 2021
Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, Philion etal 2020
BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, Liu etal 2023

Object detectors

Trend 1

- **Transformer (cross-attention) based architectures dominate**
  - End-to-end trainable architectures make up the vast majority of current detectors

Image Classification

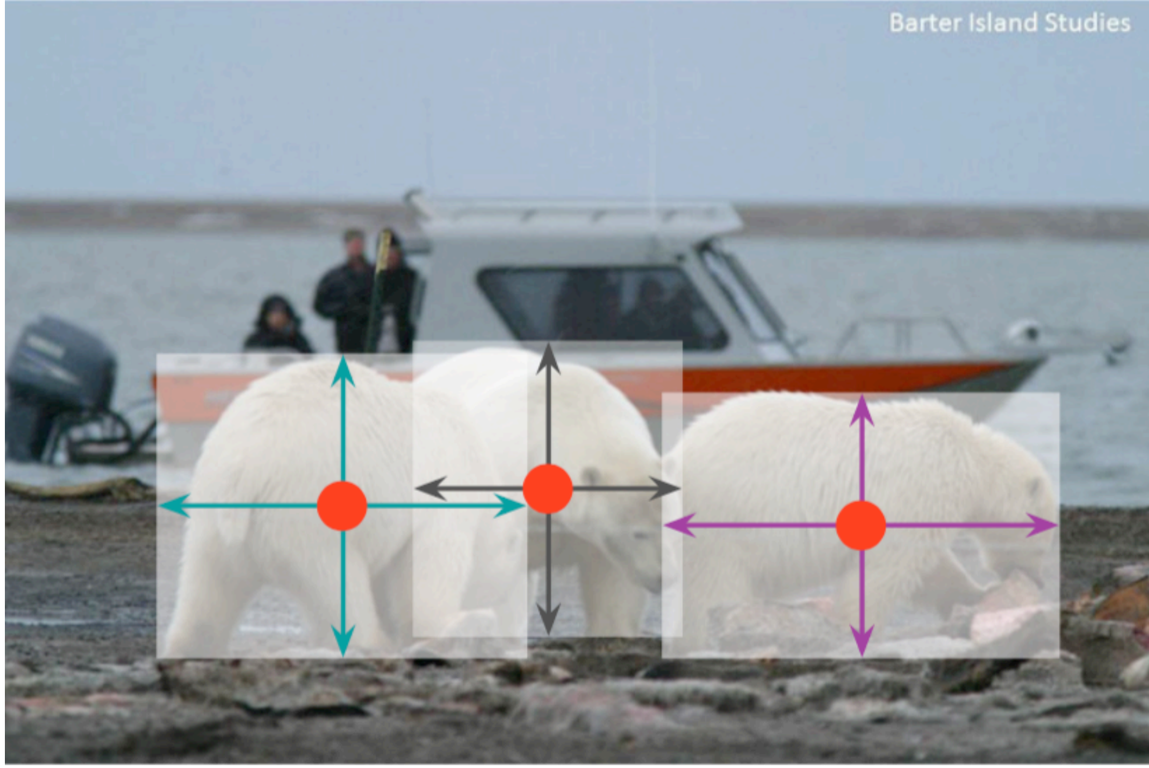
Trend 2

- **Finding objects is easy, naming them is harder**
  - Detection datasets can be smaller than classification or captioning

Object detectors

Trend 3

- **Technology is mature**
  - Used in production
    - In vehicles, photo search, surveillance
  - With **sufficient data**, detection has good solutions
  - Loosing popularity



References

- [1] Microsoft coco: Common objects in context, Lin et al., ECCV 2014
- [2] Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al., 2014
- [3] Faster R-CNN: Towards real-time object detection with region proposal networks, Ren et al., 2017
- [4] You only look once: Unified, real-time object detection, Redmon et al., 2016
- [5] Objects as Points, Xingyi Zhou, Dequan Wang, Philipp Krähenbühl, 2019
- [6] End-to-end object detection with transformers, Carion et al., 2020
- [7] Center-based 3D Object Detection and Tracking, Yin etal 2021
- [8] Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, Philion etal 2020
- [9] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, Liu etal 2023