Object detection and recognition

Cengiz Öztireli

UNIVERSITY OF CAMBRIDGE

Object Detection

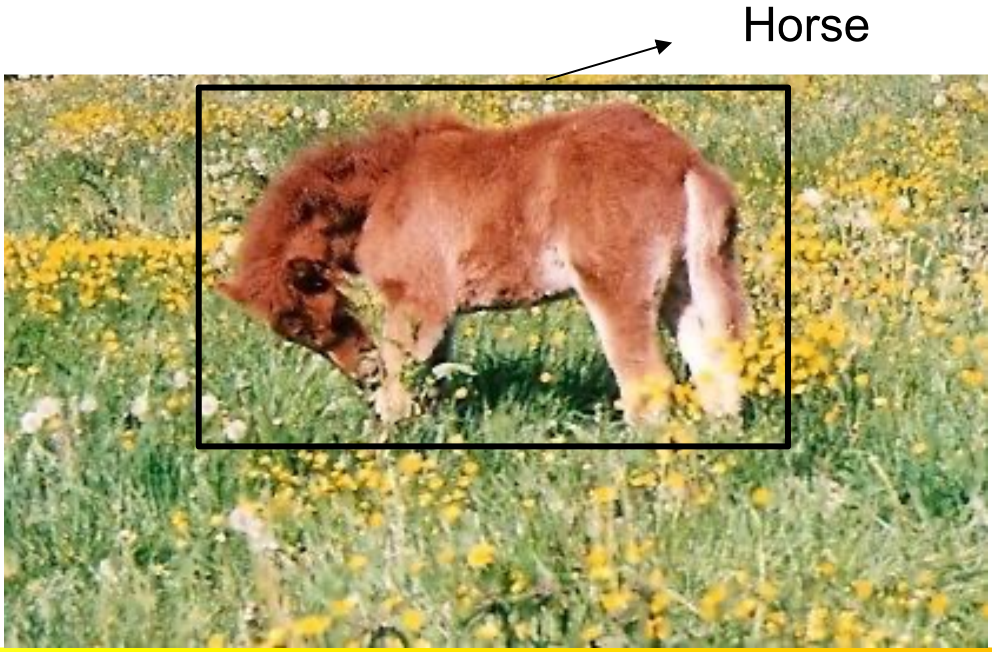
The task of assigning a label and a bounding box to all predefined objects in the image

- Localization
  - Use a bounding box to localize the objects of interests

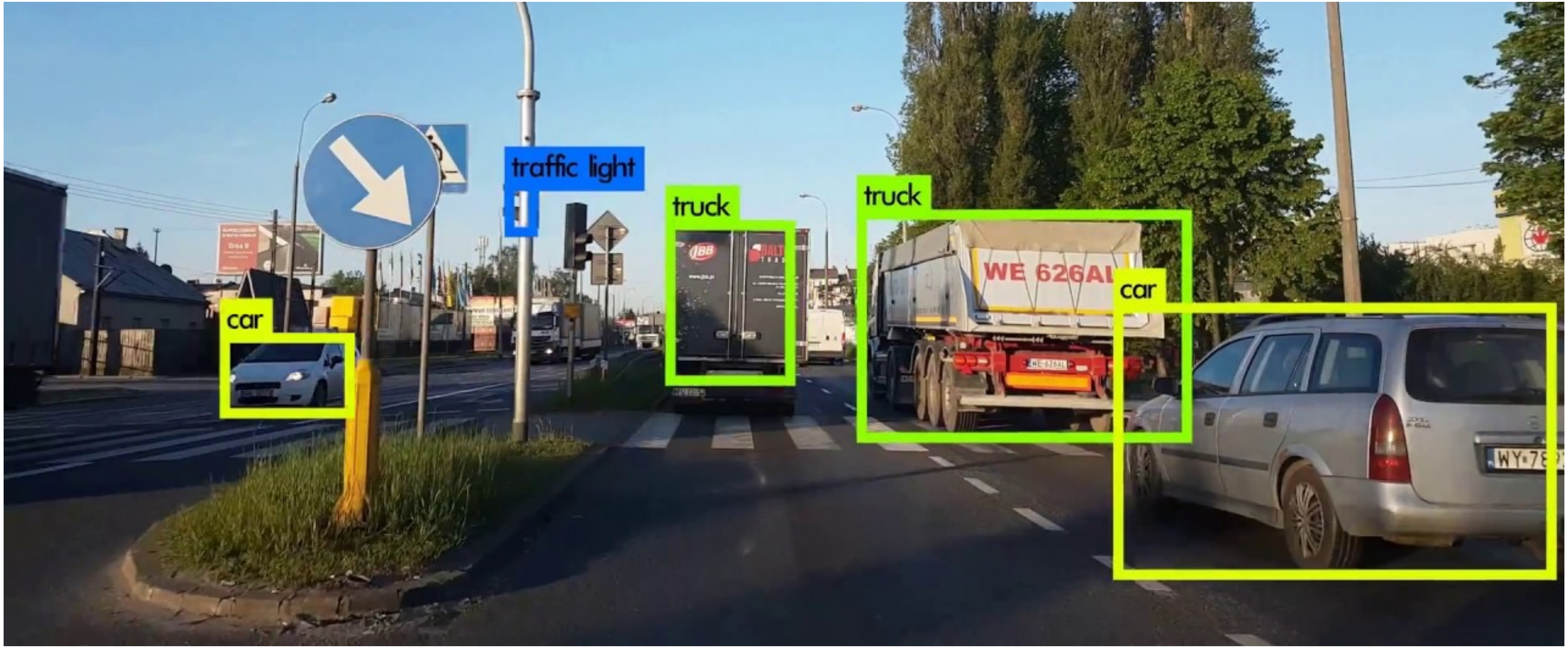
Object Detection

The task of assigning a label and a bounding box to all predefined objects in the image.

- Classification
  - Box-level classification



Applications

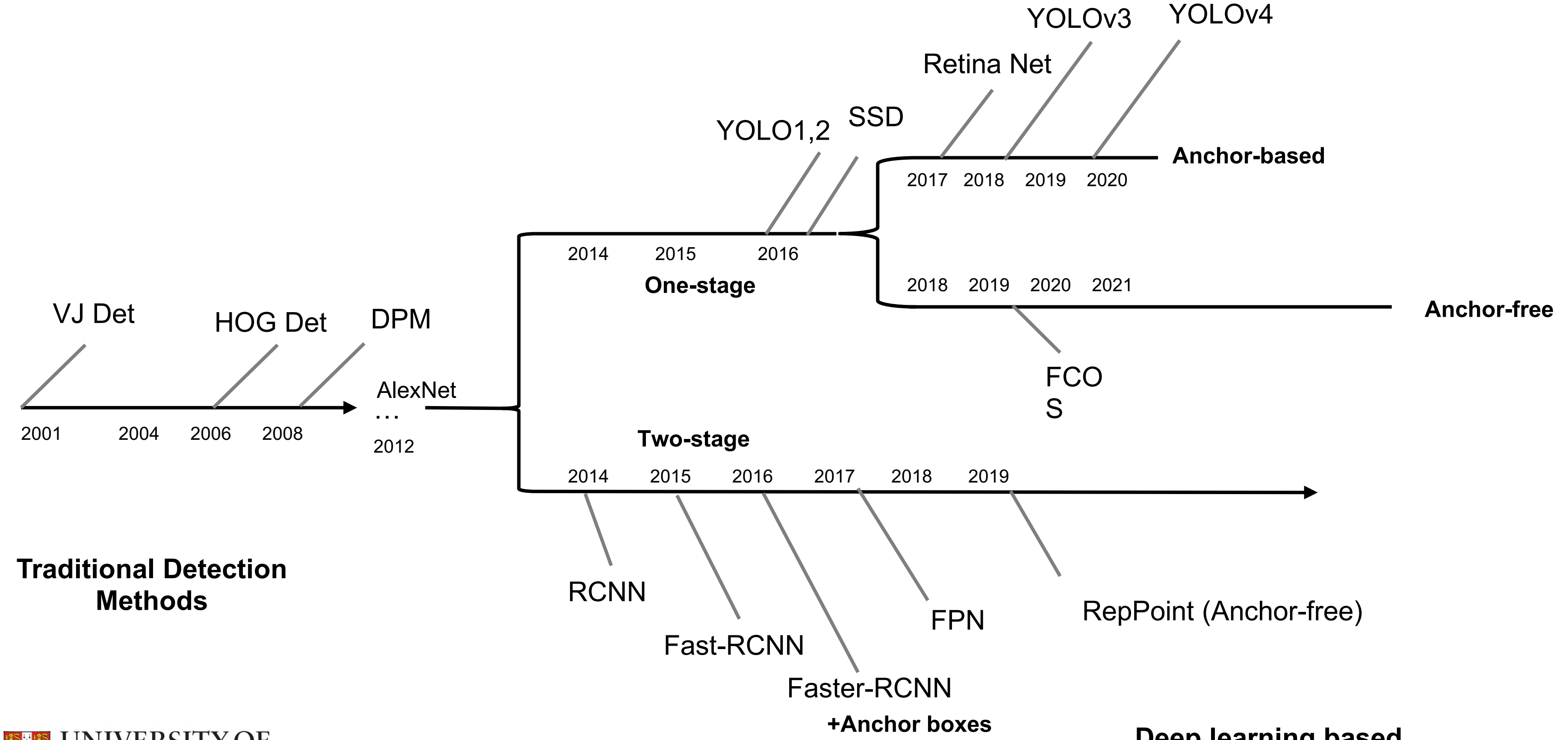


Tian, Zhi, et al. "FCOS: A Simple and Strong Anchor-free Object Detector." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 2020.

Applications

Counting crowd

Timeline



Deep learning based detection methods

UNIVERSITY OF CAMBRIDGE

Traditional Methods

Sliding windows: Classification

Class: Dog, Cat, Human, Background

Dog: Yes

Cat : No

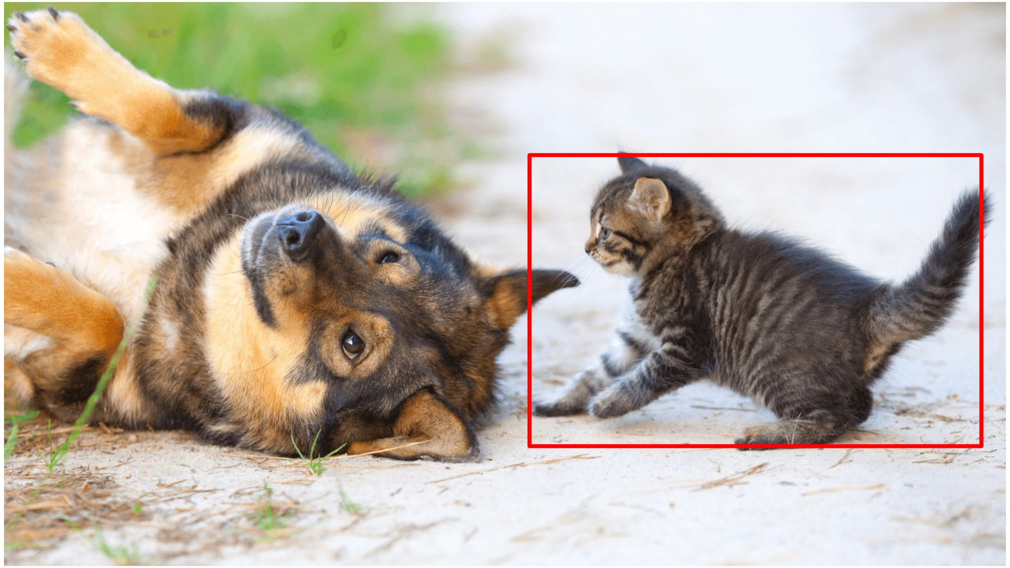
Human: No

Background: No

Traditional Methods

Sliding windows: Classification

Class: Dog, Cat, Human, Background



Dog: No

Cat : Yes

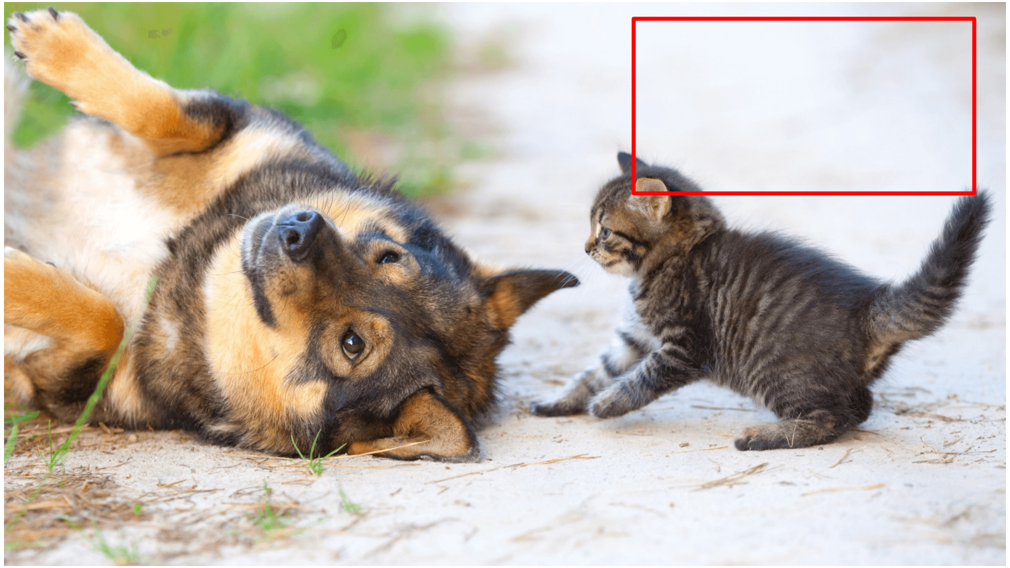
Human: No

Background: No

Traditional Methods

Sliding windows: Classification

Class: Dog, Cat, Human, Background



Dog: No

Cat : No

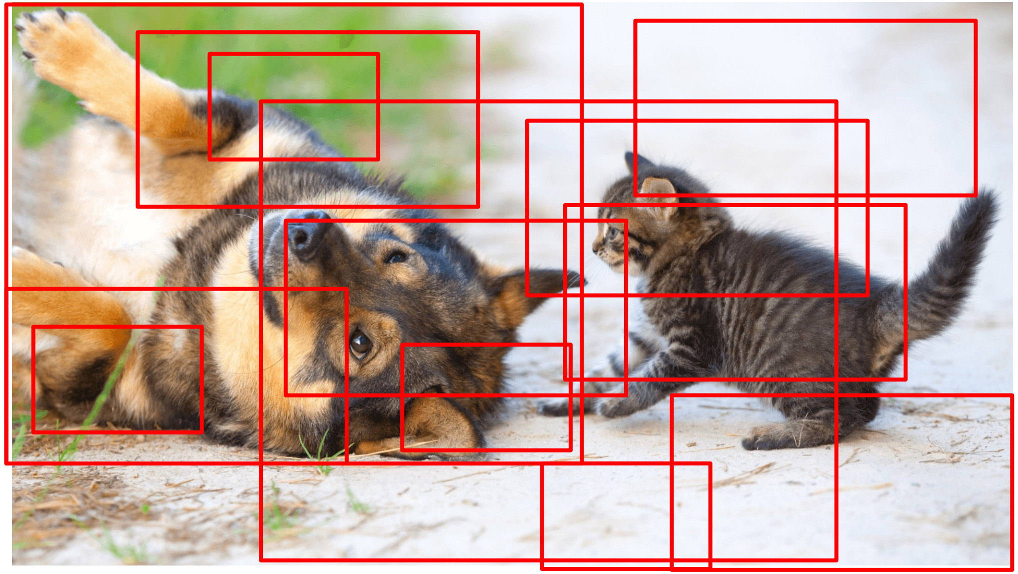
Human: No

Background: Yes

Traditional Methods

Sliding windows: Classification

Class: Dog, Cat, Human, Background



1. Compute features on multiple resolutions

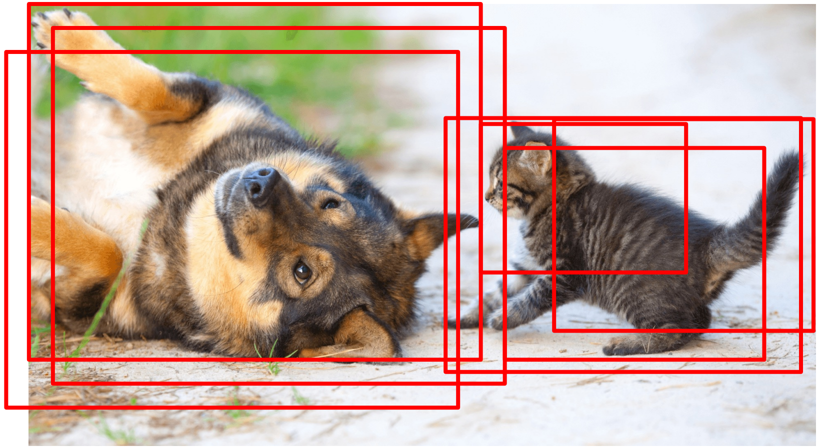
2. Scoring every sliding windows

3. Applying Non-maxima suppression

Dalal and Triggs. Histograms of Oriented Gradients for Human Detection. CVPR 2005

Non-maxima Suppression (NMS)

- **Input**: A list of proposal boxes B, corresponding confidence scores S and overlap threshold N.
- **Output**: A list of filtered proposals D.



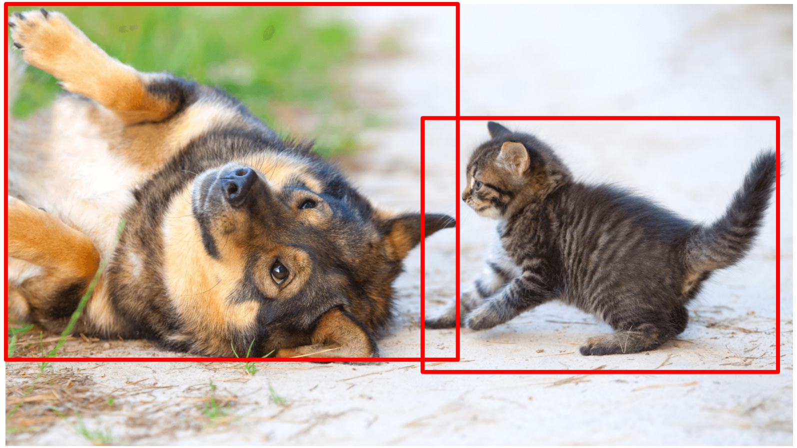
Before NMS

```
D = []
while B is not empty:
    i = Argmax(S)
    D.append[Bi]
    B.delete[Bi]

    for Br in B:
        if iou[Br,Bi] > th:
            B.delete[Br]
Return D
```

Non-maxima Suppression (NMS)

- **Input**: A list of proposal boxes B, corresponding confidence scores S and overlap threshold N.
- **Output**: A list of filtered proposals D.



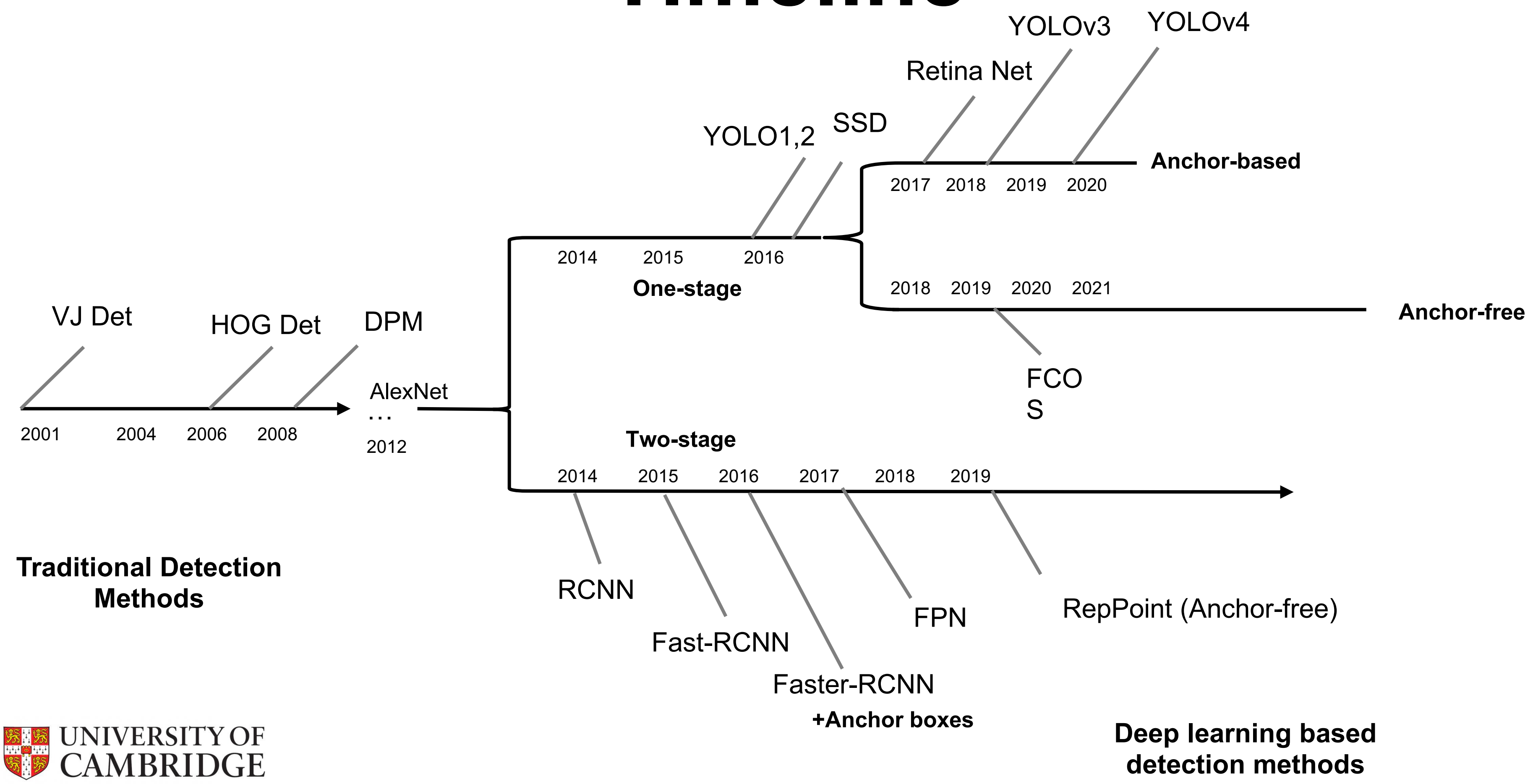
After NMS

```
D = []
while B is not empty:
    i = Argmax(S)
    D.append[Bi]
    B.delete[Bi]

    for Br in B:
        if iou[Br,Bi] > th:
            B.delete[Br]
Return D
```

UNIVERSITY OF CAMBRIDGE

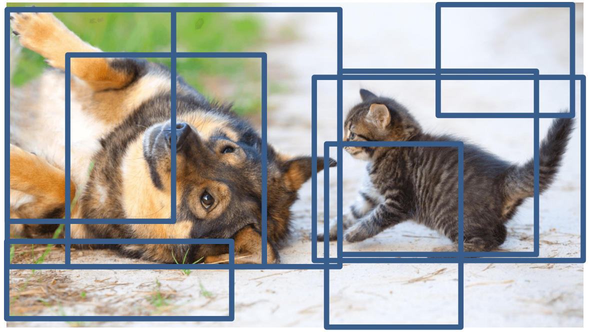
Timeline



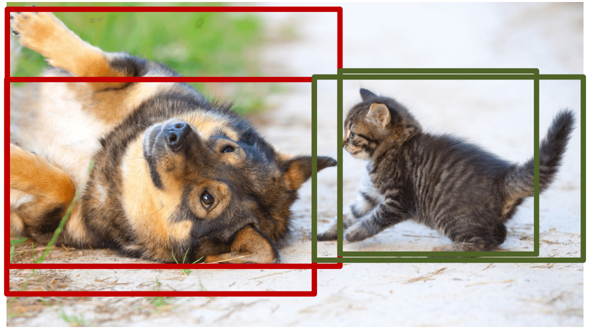
YOLOv3
YOLOv4
Retina Net
YOLO1,2
SSD
Anchor-based
2017 2018 2019 2020
2014 2015 2016
One-stage
2018 2019 2020 2021
Anchor-free
VJ Det
HOG Det
DPM
FCO
S
AlexNet
…
2001 2004 2006 2008
2012
Two-stage
2014 2015 2016 2017 2018 2019
Traditional Detection
Methods
RCNN
Fast-RCNN
Faster-RCNN
+Anchor boxes
FPN
RepPoint (Anchor-free)
Deep learning based
detection methods

Two stage vs. One stage

Input image



Proposals

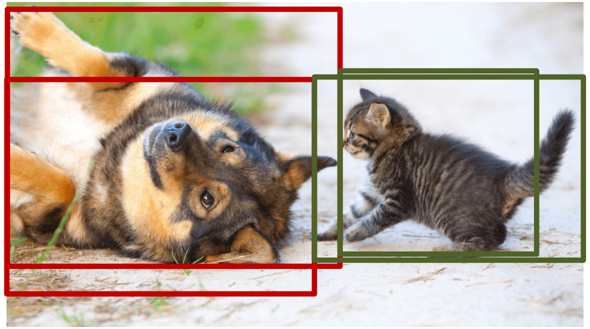


Classify the proposals

Two stage methods

Two stage vs. One stage

Input image



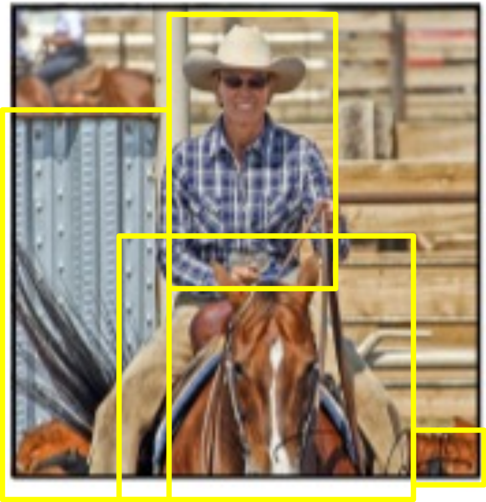
Predict classified boxes

One stage methods

RCNN

Step 1: Selective Search → ~2k proposals

- find image regions that likely contain objects



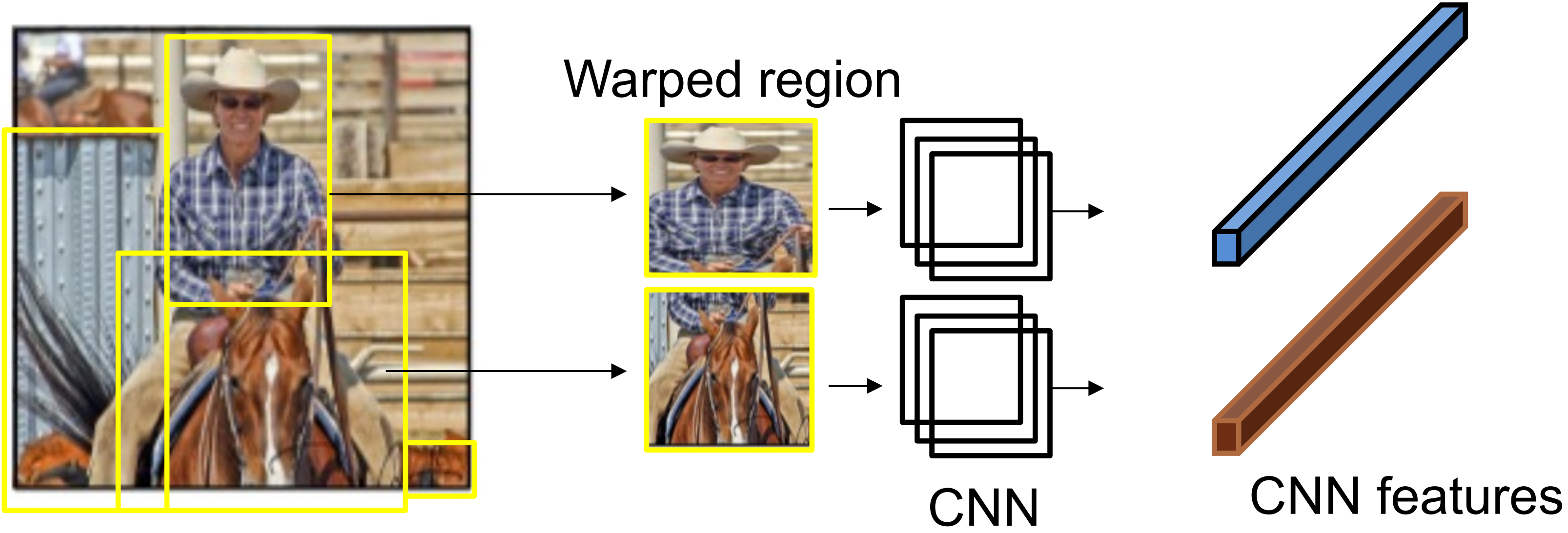
UNIVERSITY OF CAMBRIDGE

Uijlings, J. R R, Sande V D, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.

RCNN

Step 2: CNN extract features

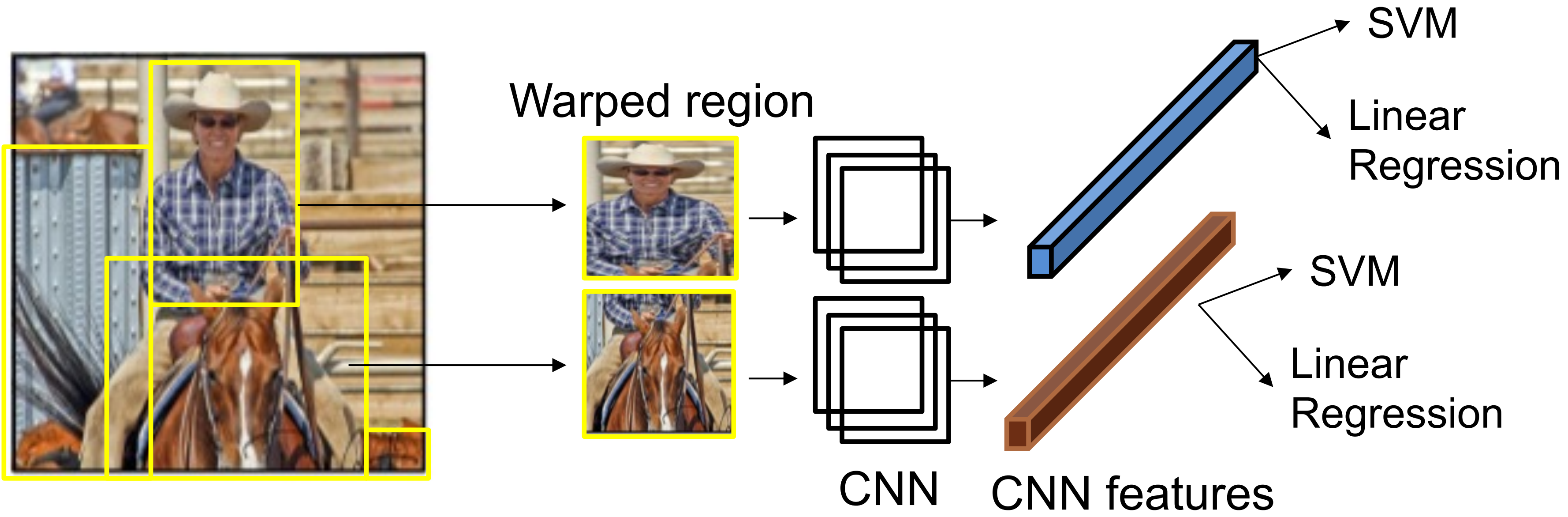
- Affine image warping: Get a fixed input size



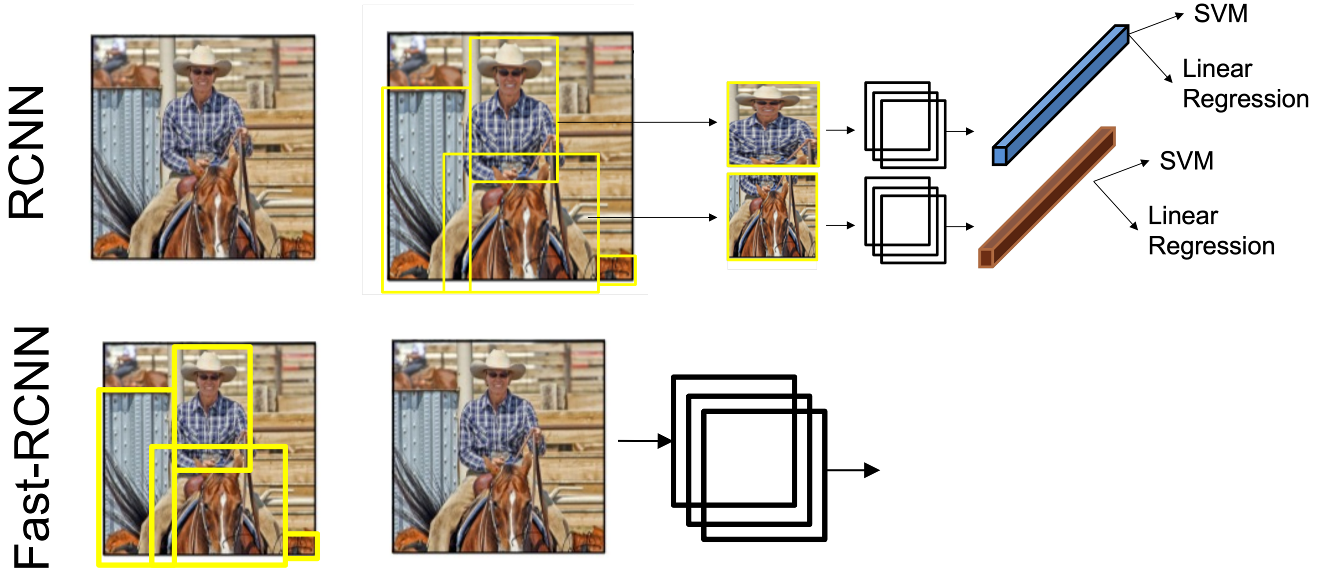
RCNN

Step 3: Classification and regression

- Classify with a linear SVM, linear regression for the bounding box offset



Fast-RCNN

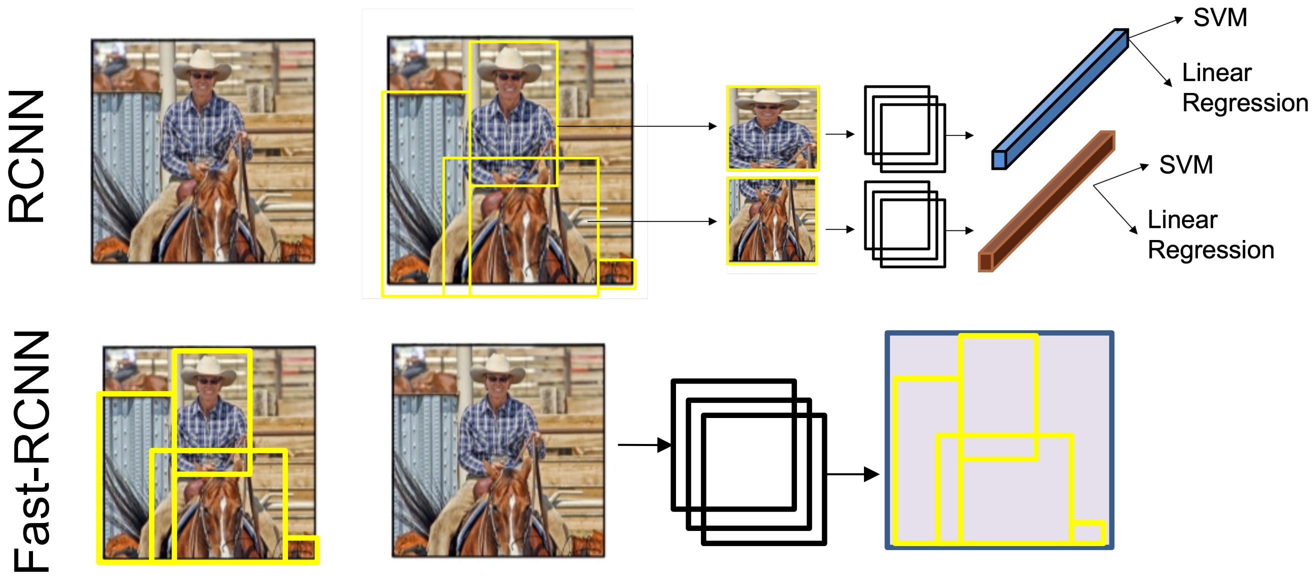


1. Selective search

2. Input image to CNN

Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

Fast-RCNN



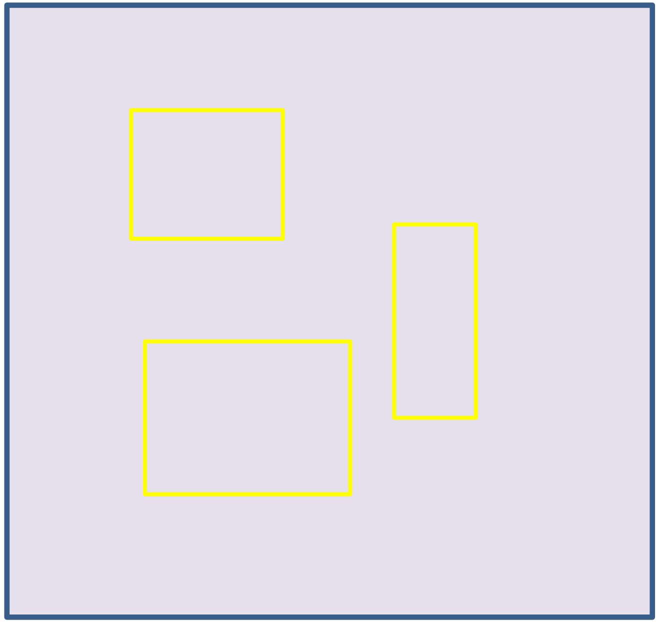
1. Selective search

3. ROI pooling

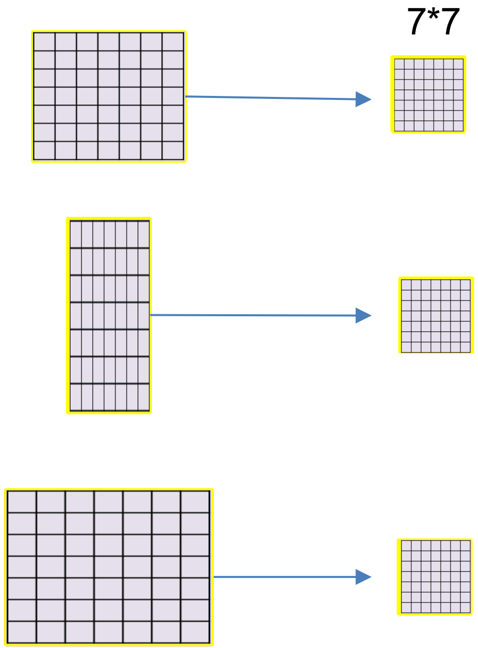
2. Input image to CNN

Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

ROI Pooling



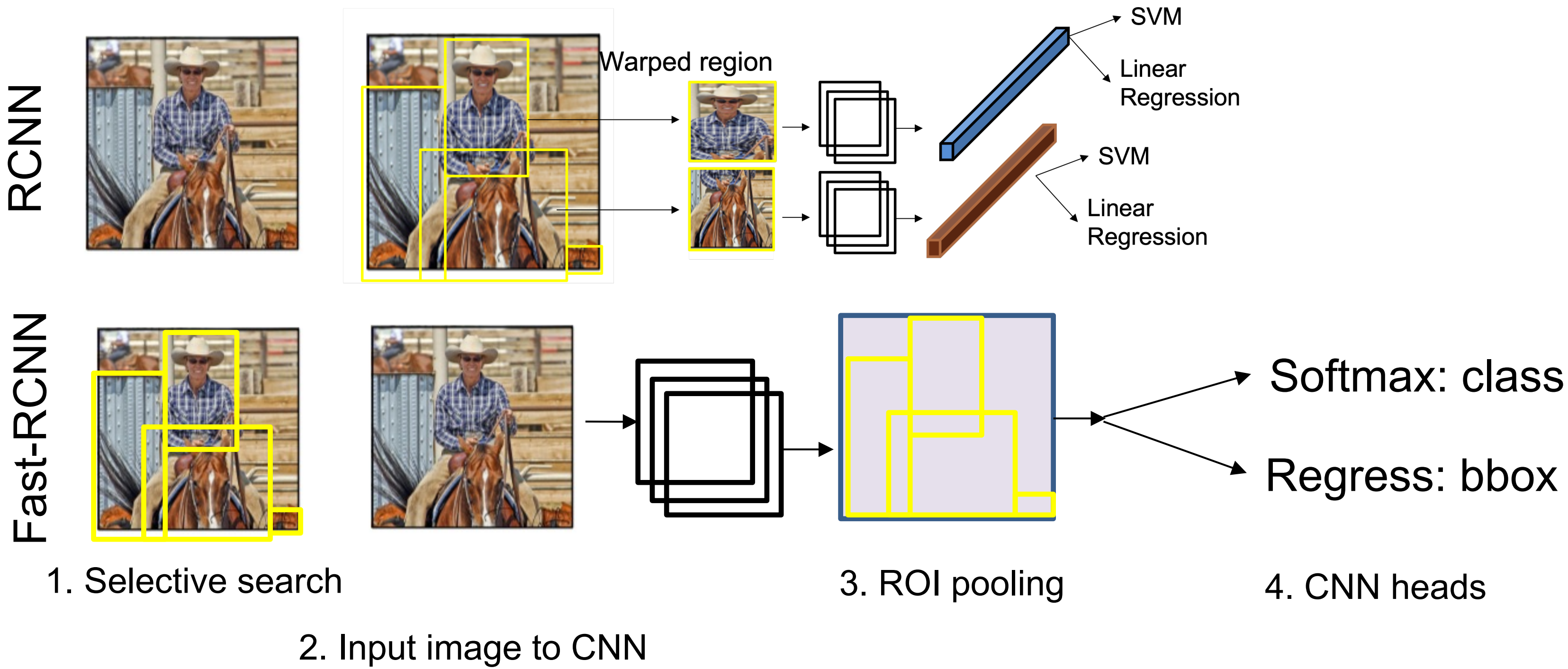
Feature map



ROI Pooling: max pooling for each grid

UNIVERSITY OF CAMBRIDGE

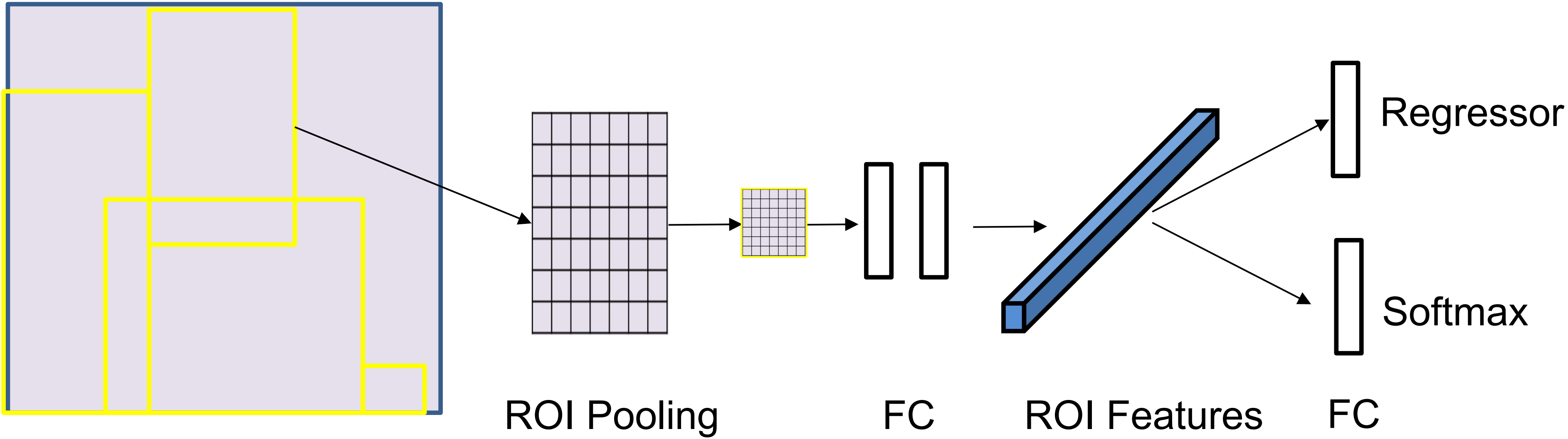
Fast-RCNN



Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

Fast-RCNN

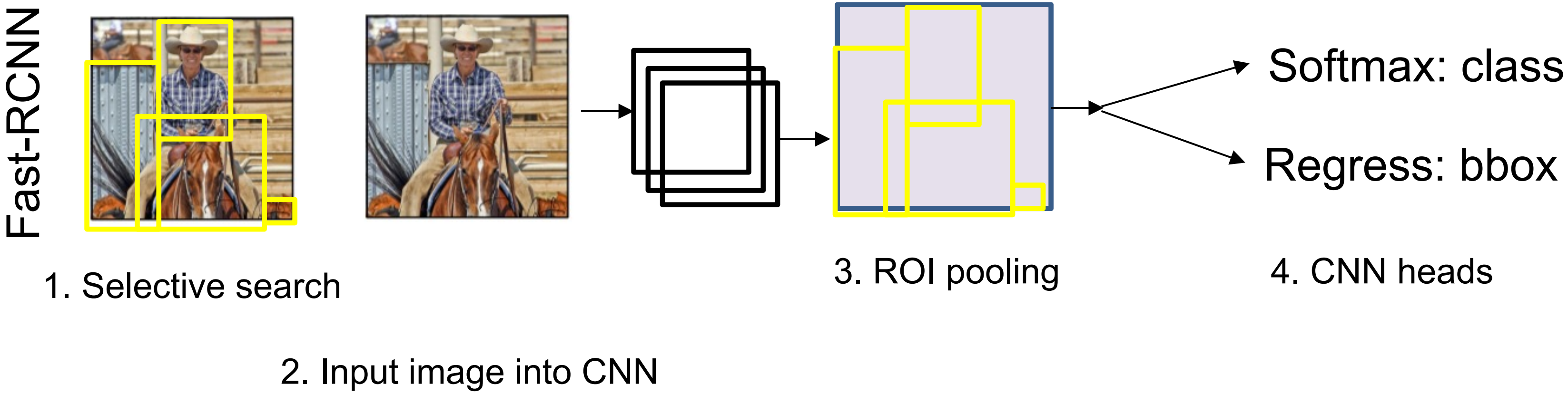
Details of the CNN head



Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

Faster-RCNN

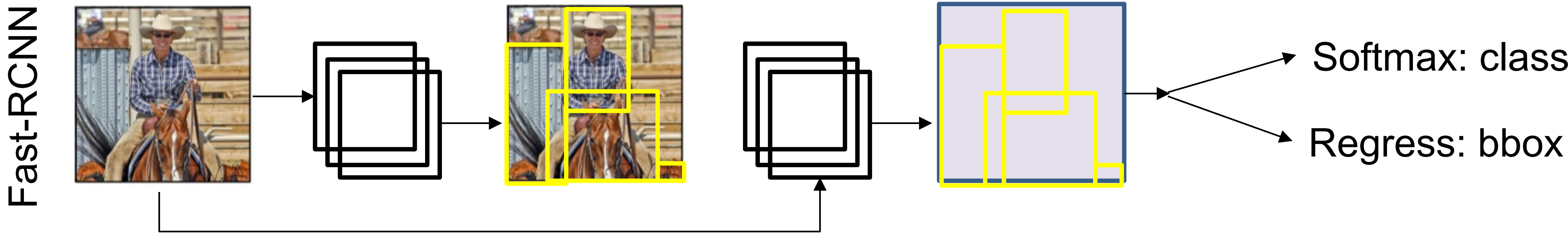
Main difference: Use a CNN to generate region proposals instead of selective search



Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015): 91-99.

Faster-RCNN

Main difference: Use a CNN to generate region proposals instead of selective search



1. Input image into CNN

3. ROI pooling

4. CNN heads

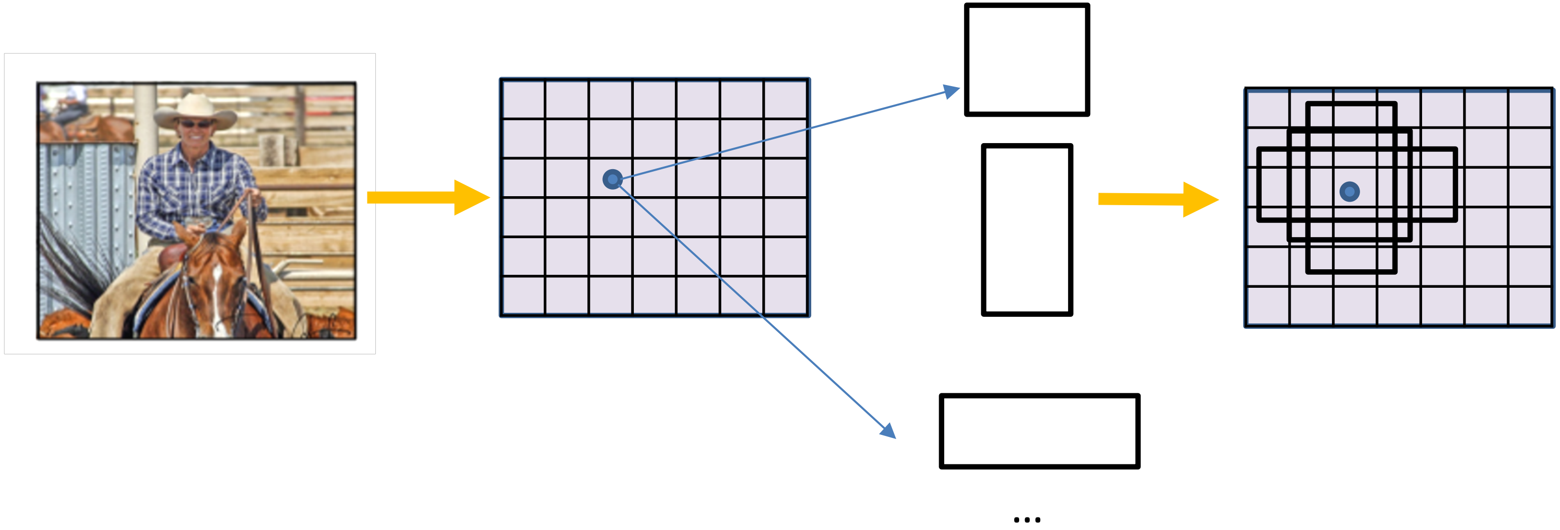
2. Region Proposal Network

RPN

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015): 91-99.

Faster-RCNN: RPN

Anchors: pre-defined boxes

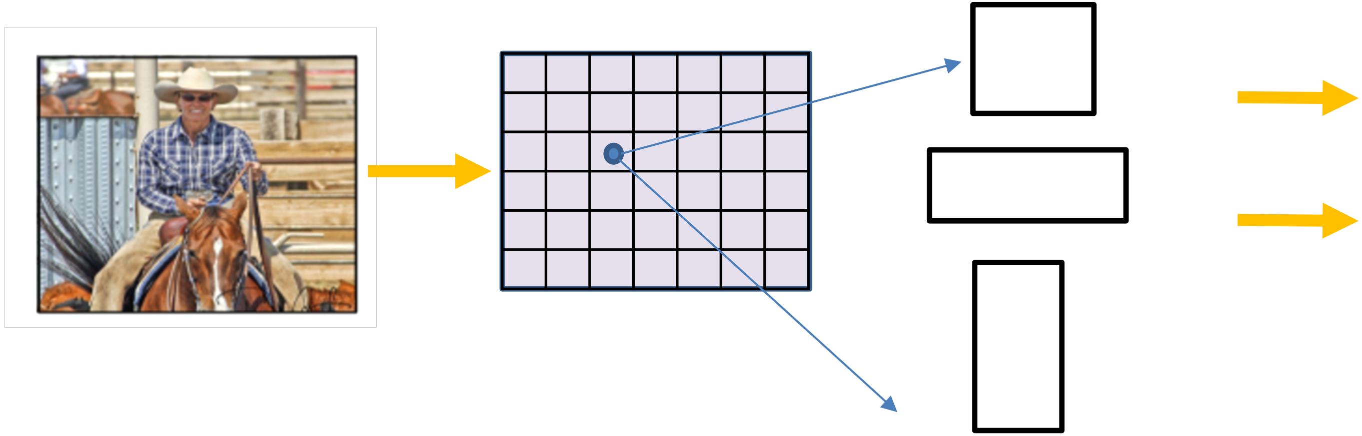


UNIVERSITY OF CAMBRIDGE

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015): 91-99.

Faster-RCNN: RPN

Anchors: pre-defined boxes



...

9 pre-defined shapes

For each anchor

Positive/Negative

Regress bbox (x,y,w,h)

Output **positive** bbox

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015): 91-99.

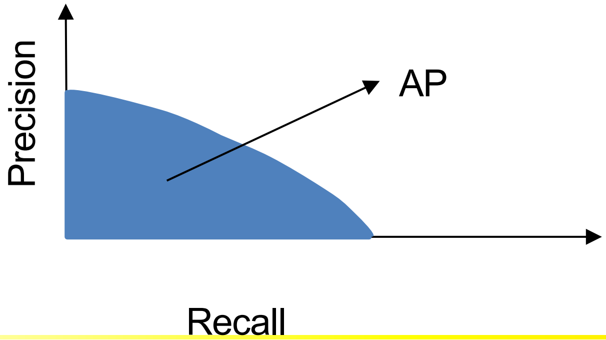
Test Performance

	RCNN	Fast-RCNN	Faster RCNN
Time	50 s	2 s	0.2 s
mAP on Pascal VOC	66.0	66.9	66.9

Mean Average Precision

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

TP = True Positives (Predicted as positive as was correct)
FN = False Negatives (Failed to predict an object)
FP = False Positives (Predicted as positive but was incorrect)



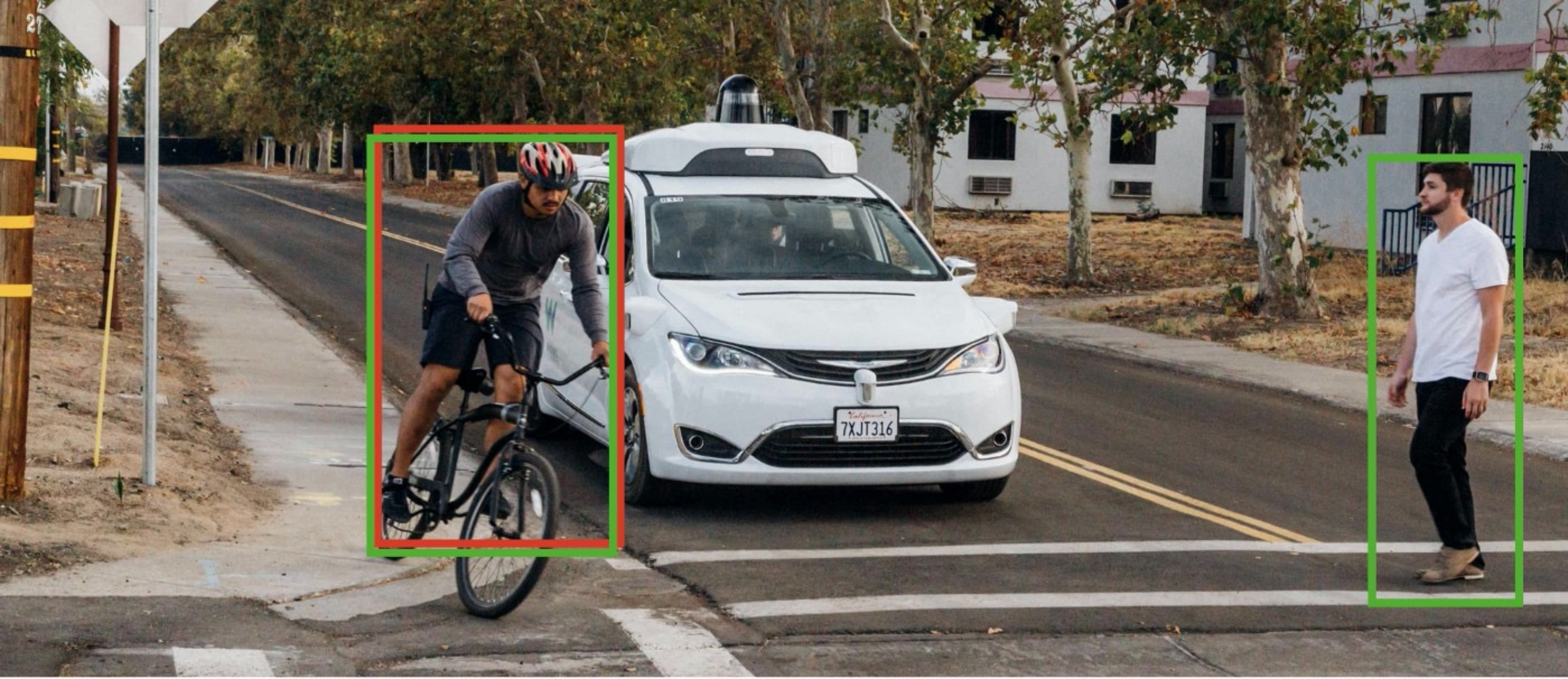


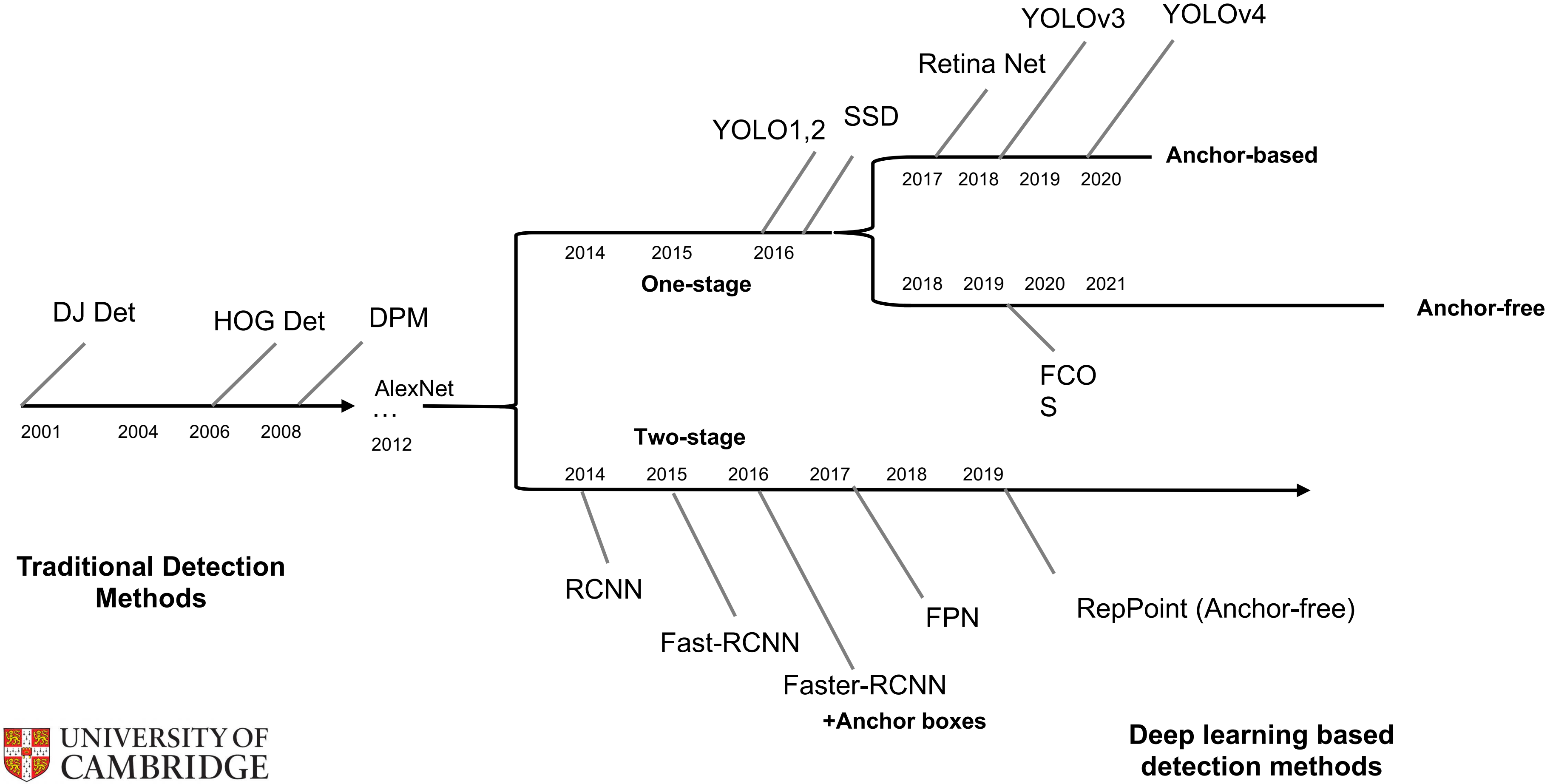
image: Waymo

= Predicted Bounding Box

= Ground Truth Bounding Box

TP=1, FP=0, FN=1

Timeline



UNIVERSITY OF CAMBRIDGE

Two-stage vs One-stage

One stage: No region proposal networks

Directly regress the bbox (dx, dy, dh, dw, confidence)

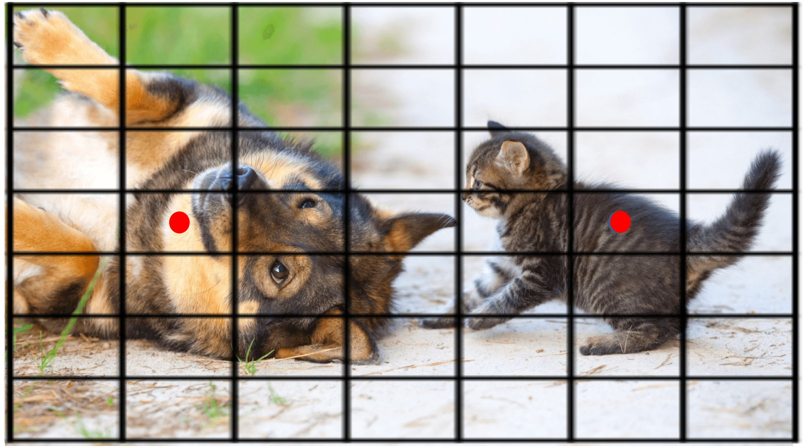
YOLO, SSD

Two-stage vs One-stage

One stage: No region proposal networks

Directly regress the bbox (dx, dy, dh, dw, confidence)

YOLO, SSD



7*7 grids

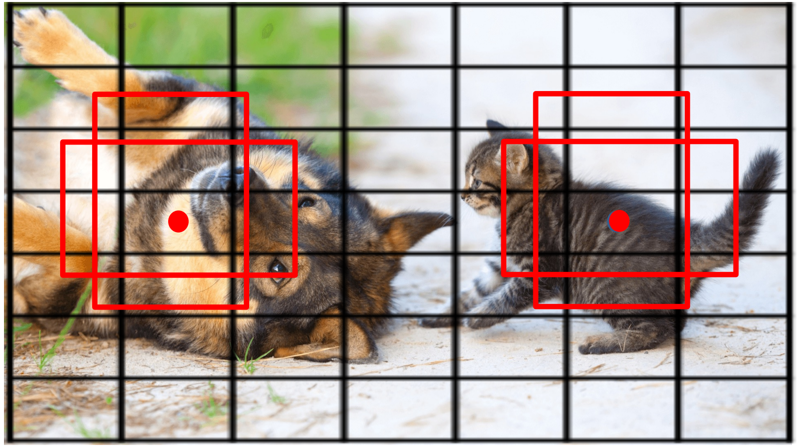
YOLO V1: only one estimation per pre-defined region → low recall

Two-stage vs One-stage

One stage: No region proposal networks

Directly regress the bbox (dx, dy, dh, dw, confidence)

YOLO, SSD



7*7 grids

YOLO V1: only one estimation per pre-defined region → low recall

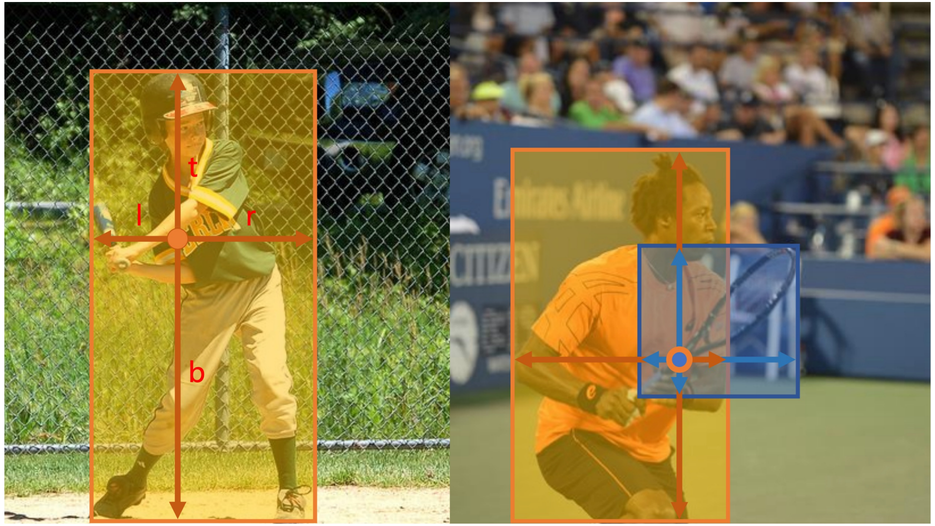
YOLOV2, 3: Use anchors → improve recall

Anchor-free vs Anchor-based

Anchor-free: do not rely on anchors

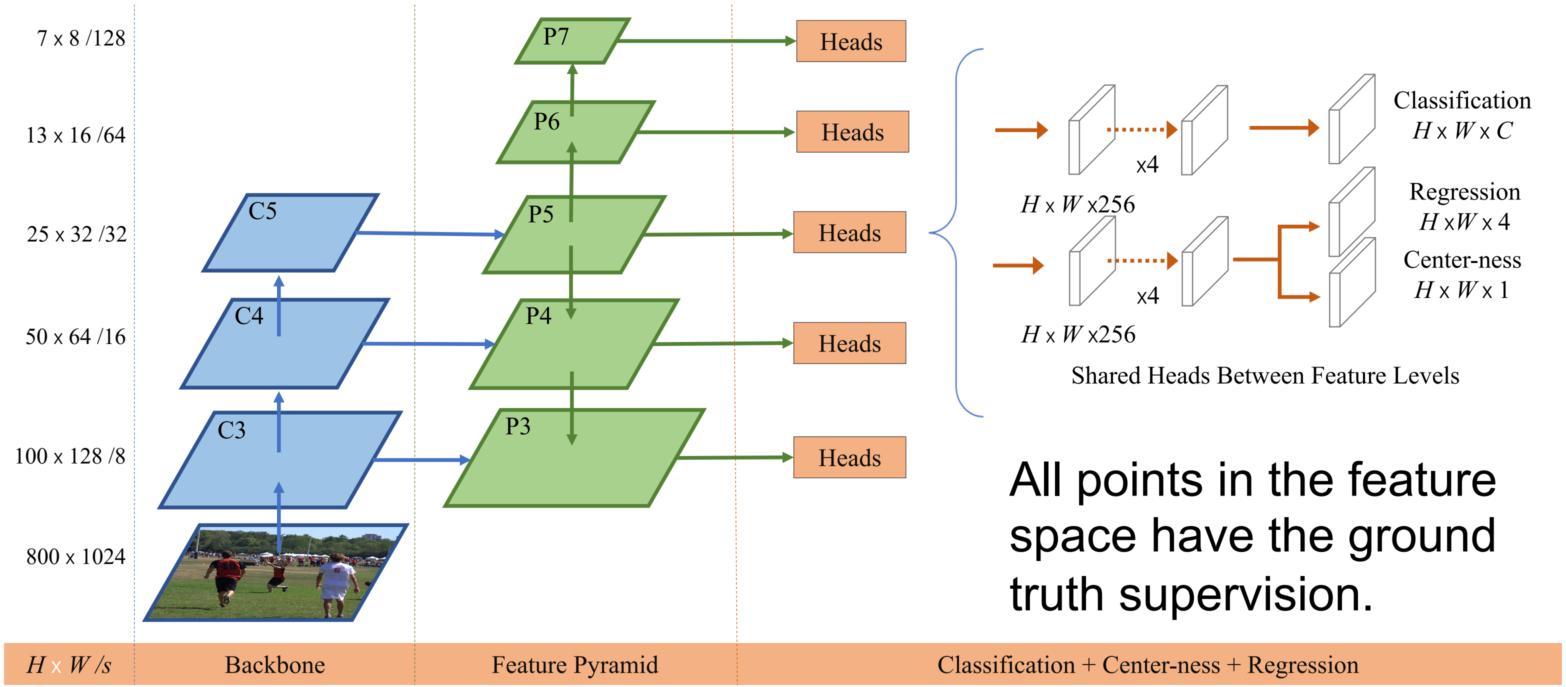
YOLOv1: only one estimation per pre-defined region

FCOS: for each point on the image plane, regress to the distances to the bounding box edges → high recall



Tian, Zhi, Chunhua Shen, Hao Chen, and Tong He. "Fcos: Fully convolutional one-stage object detection." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9627-9636. 2019.

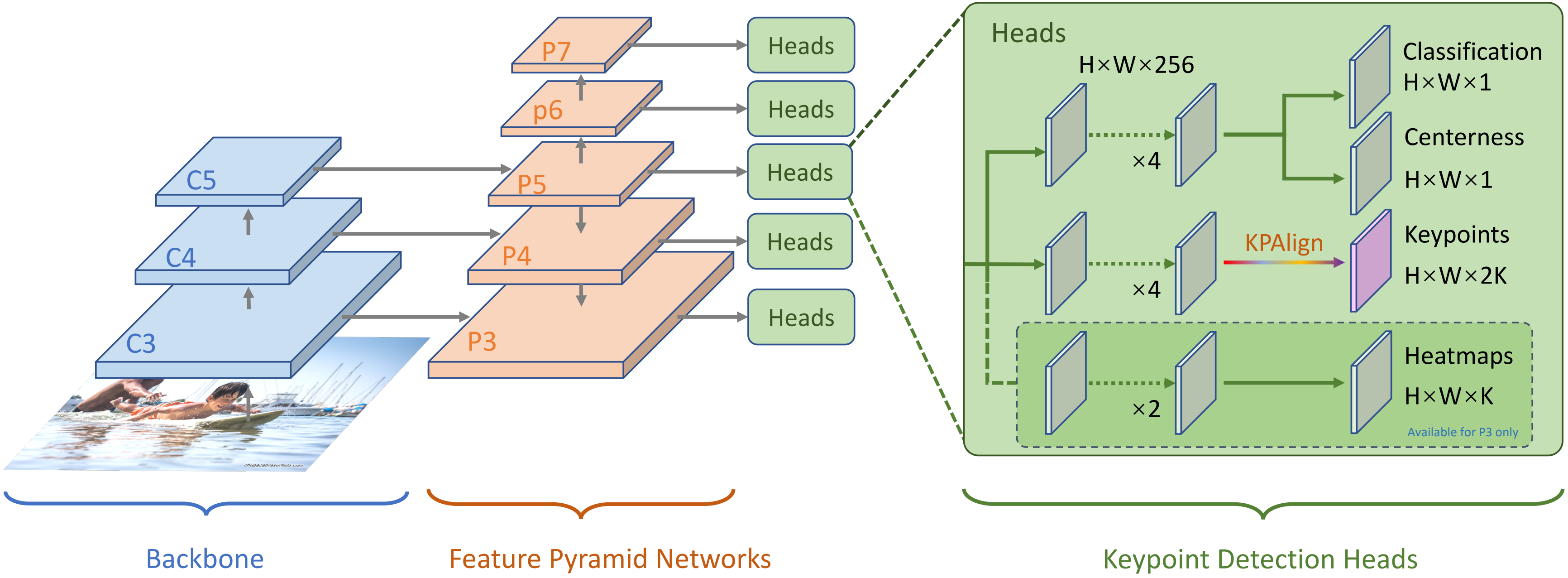
FCOS



All points in the feature space have the ground truth supervision.

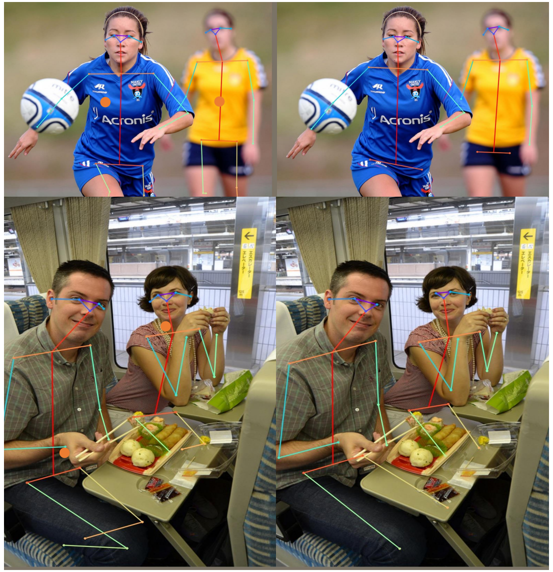
UNIVERSITY OF CAMBRIDGE

Aside: Key-point Detection



Tian, Zhi, Hao Chen, and Chunhua Shen. "Directpose: Direct end-to-end multi-person pose estimation." arXiv preprint arXiv:1911.07451 (2019).

Aside: Keypoint Detection



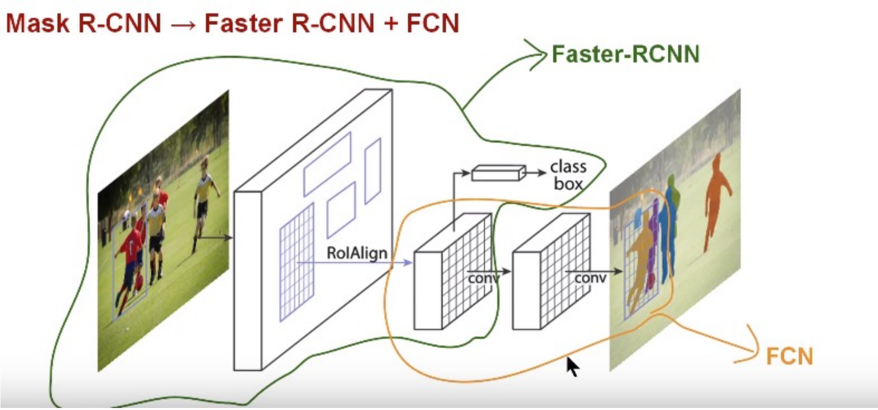
Estimated Ground Truth

Estimated Ground Truth

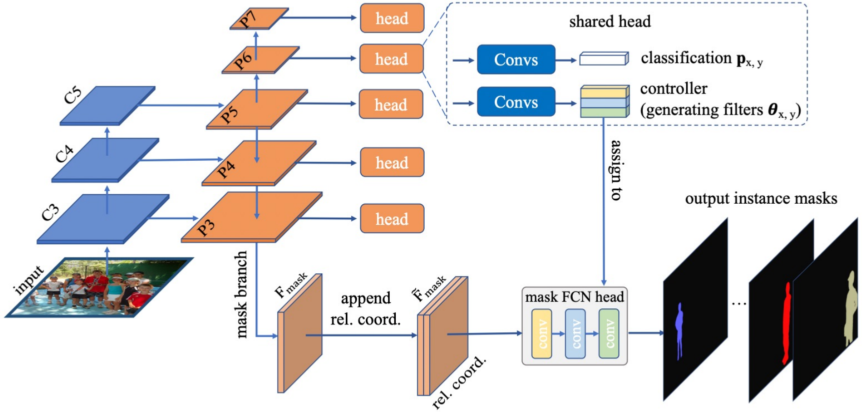
UNIVERSITY OF CAMBRIDGE

Aside: Instance Segmentation

Aside: Instance Segmentation

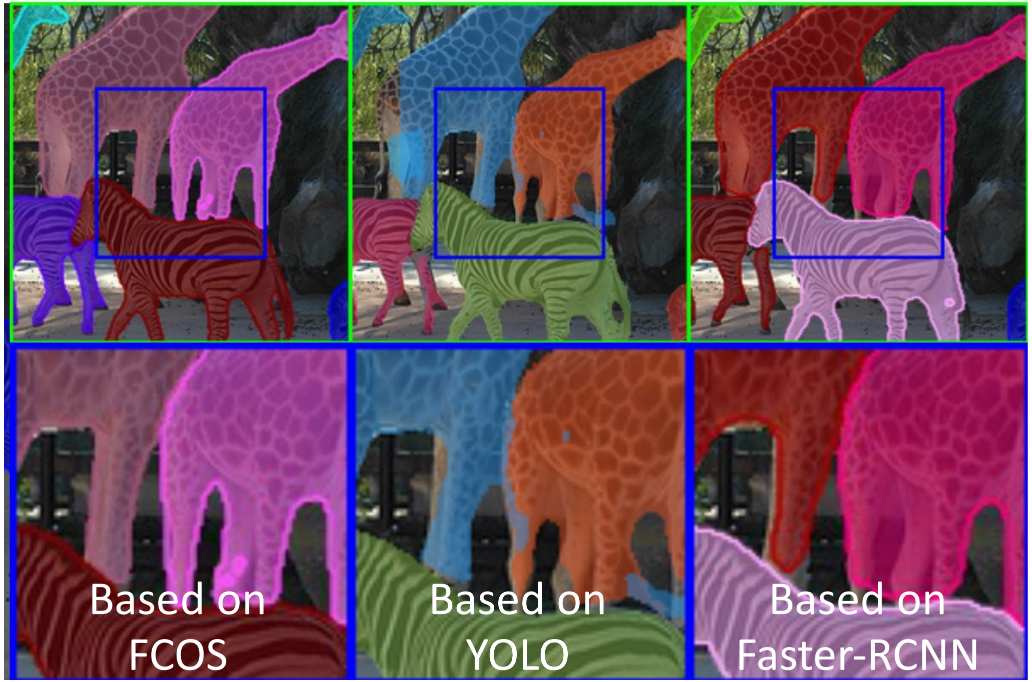


Based on Faster-RCNN



Based on FCOS

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.


Based on
FCOS
Based on
YOLO
Based on
Faster-RCNN

Multiple Variations

- **Input**: Image, Patch, Image Pyramid
- **Backbones:**
  - VGG16, ResNet-50, SpineNet, EfficientNet-B0/B7, CSPResNeXt50, CSPDarknet53
- **Neck:**
  - **Additional block:** SPP, ASPP, RFB, SAM
  - **Feature Fusion:** FPN, PAN, NAS-FPN, Fully-connected FPN, BiFPN, ASFF, SFAM
- **Head:**
  - **One-stage:**
    - RPN, SSD, YOLO, RetinaNet (anchor based)
    - CornerNet, CenterNet , MatrixNet , FCOS (anchor free)
  - **Two-stage:**
    - Faster R-CNN, R-FCN, Mask RCNN (anchor based)
    - RepPoints (anchor free)

New Framework based on Transformers

- Examples:
  - DETR: End-to-End Object Detection with Transformers
  - Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

COCO dataset

	FasterRCNN	FasterRCNN-fpn	FCOS	Retinanet	DETR	SWIN-L
mAP	41.1	42.0	43.1	40.4	44.9	58

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers." In European Conference on Computer Vision, pp. 213-229. Springer, Cham, 2020.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).

Reference List

1) Viola, Paul, and Michael Jones. "Robust real-time object detection." International journal of computer vision 4, no. 34-47 (2001): 4.
2) Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886-893. Ieee, 2005.
3) Felzenszwalb, Pedro F., Ross B. Girshick, David McAllester, and Deva Ramanan. "Object detection with discriminatively trained part-based models." IEEE transactions on pattern analysis and machine intelligence 32, no. 9 (2009): 1627-1645.
4) He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." IEEE transactions on pattern analysis and machine intelligence 37, no. 9 (2015): 1904-1916.
5) Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
6) Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.
7) Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015): 91-99.
8) Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125. 2017.
9) Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection." In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988. 2017.
10) Yang, Ze, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. "Reppoints: Point set representation for object detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9657-9666. 2019.
11) Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.
12) Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271. 2017.
13) Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).
14) Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).