Object detection

Vilde Schulerud Bøe

Slides based on Dhananjay Tomar

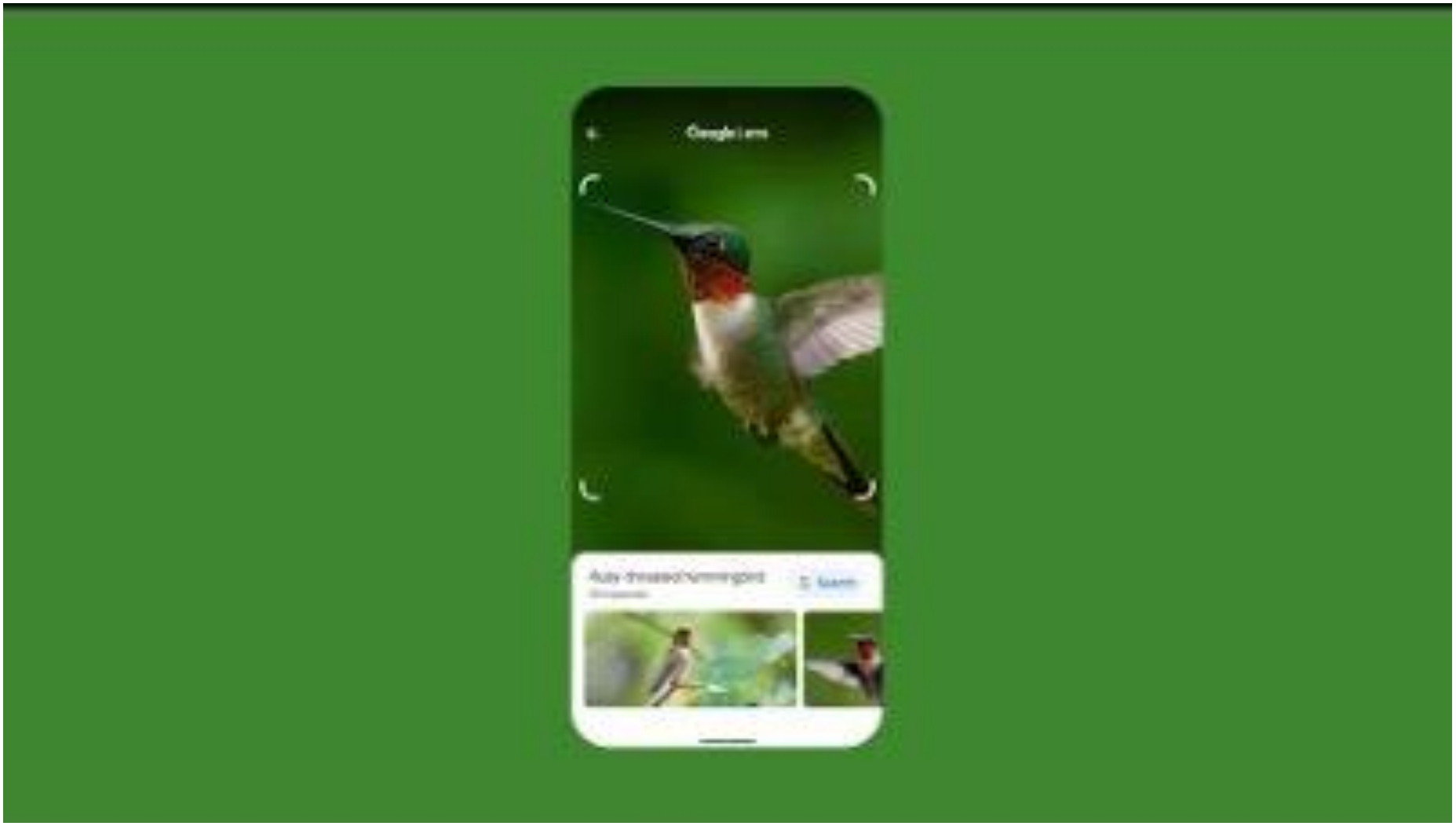
IN4310/3310

24th April 2026

- **Single-instance detection and localization** ← Know enough to implement it
- **Object detection (multiple objects)** ← Know core principles and techniques
- **Performance metrics** ← Know the most common ones



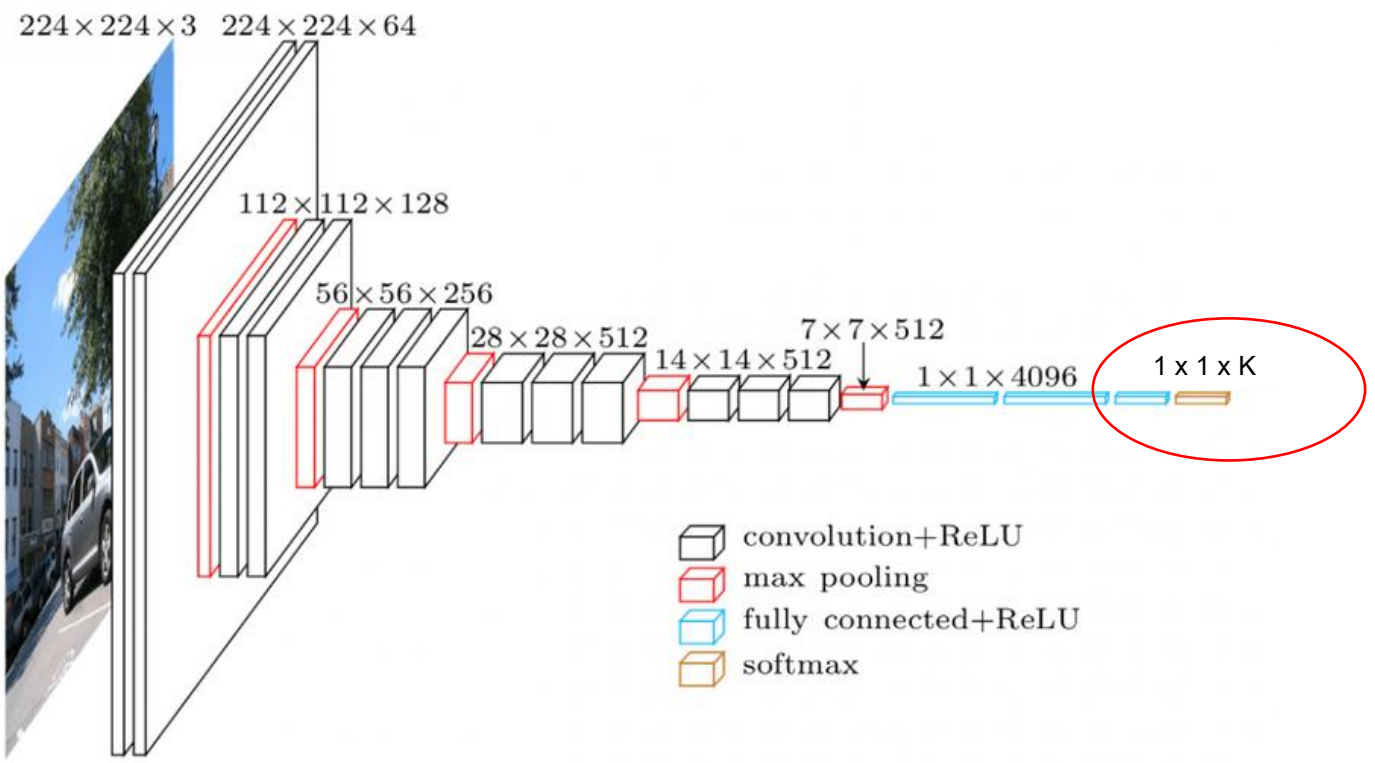
Image source: [Pinterest's blog post on medium]



https://www.youtube.com/watch?v=BH_6BdgTdiw

You already know this!

Image classification



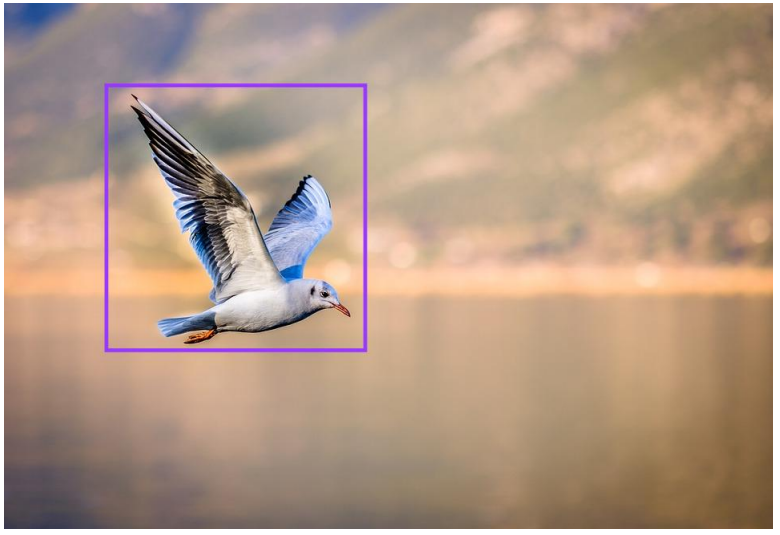
Network output

$$\mathbf{y} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_K \end{bmatrix}$$

The VGG16 network; however, think of it as just a conceptual illustration of a convolutional network

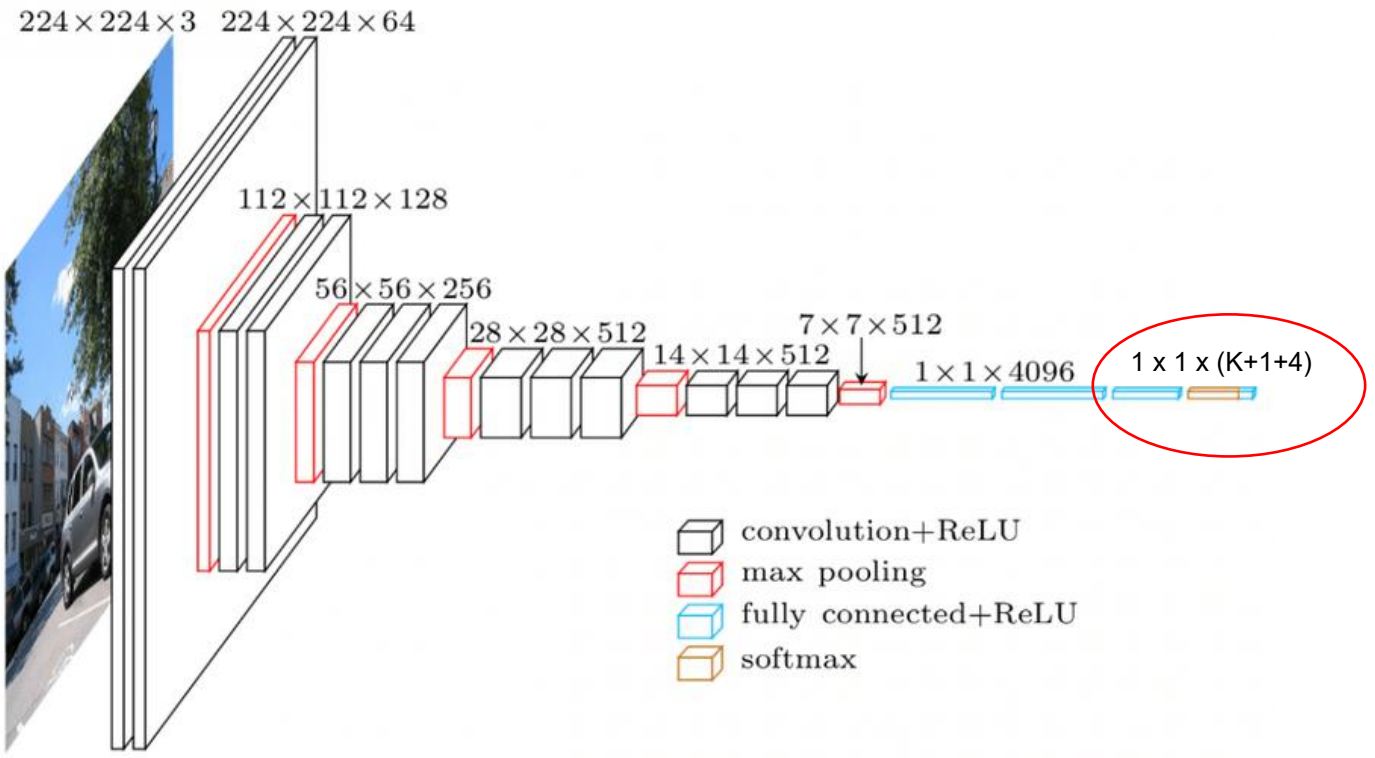
Image classification and object localization

- *Classify* an image with (at most) *a single* object
- Draw a *bounding box* around the object



b_r: Center row coordinate
b_c: Center column coordinate
b_h: Box height
b_w: Box width

Image classification + localization



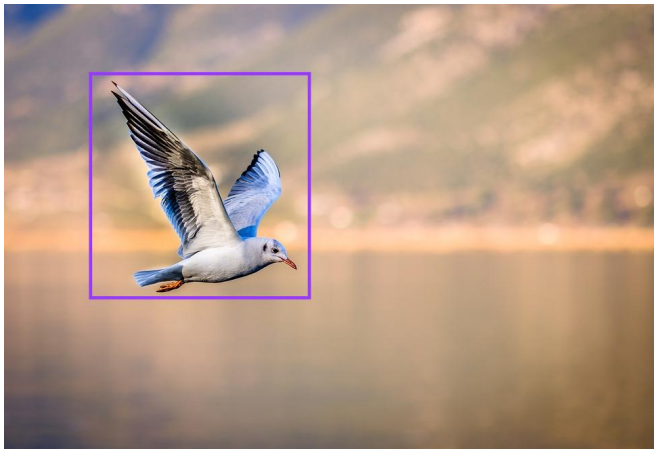
$$\mathbf{y} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_K \\ b_r \\ b_c \\ b_h \\ b_w \end{bmatrix}$$

Background class / no object

$$\mathbf{y} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_K \\ b_r \\ b_c \\ b_h \\ b_w \end{bmatrix}$$

Class probabilities

Bounding box



b_r: Center row coordinate
b_c: Center column coordinate
b_h: Box height
b_w: Box width

Example I/III

Note: These examples show what our network ideally should have as output for such images, cf. training data.

c_1: Tiger

c_2: Leopard

c_3: Lion

$$y = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ b_r \\ b_c \\ b_h \\ b_w \end{bmatrix}$$



Figure 17: Tiger. Image source: https://www.pixabay.com

Example II/III

c_1: Tiger

c_2: Leopard

c_3: Lion

$$y = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ b_r \\ b_c \\ b_h \\ b_w \end{bmatrix}$$

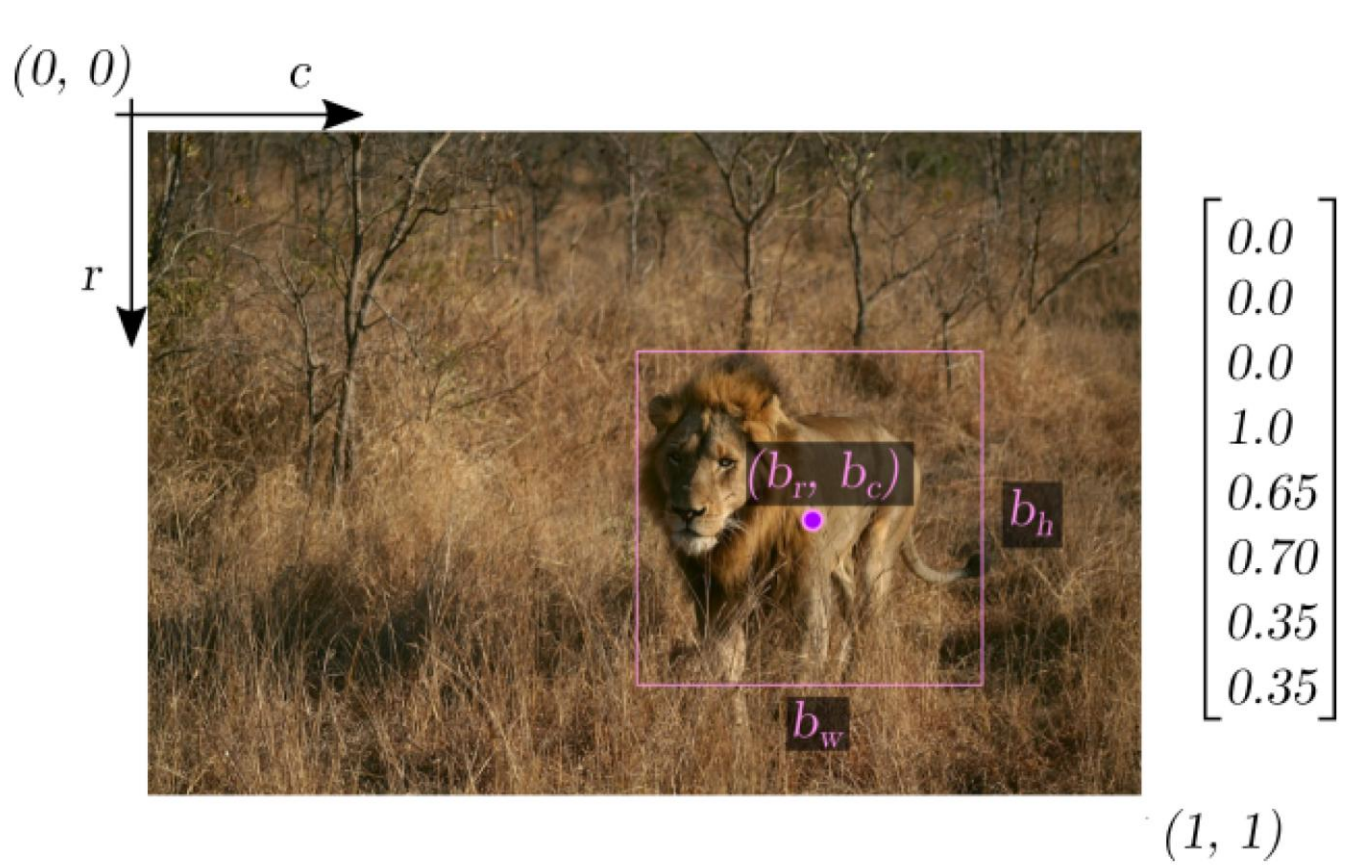


Figure 18: Lion. Image source: https://www.pixabay.com

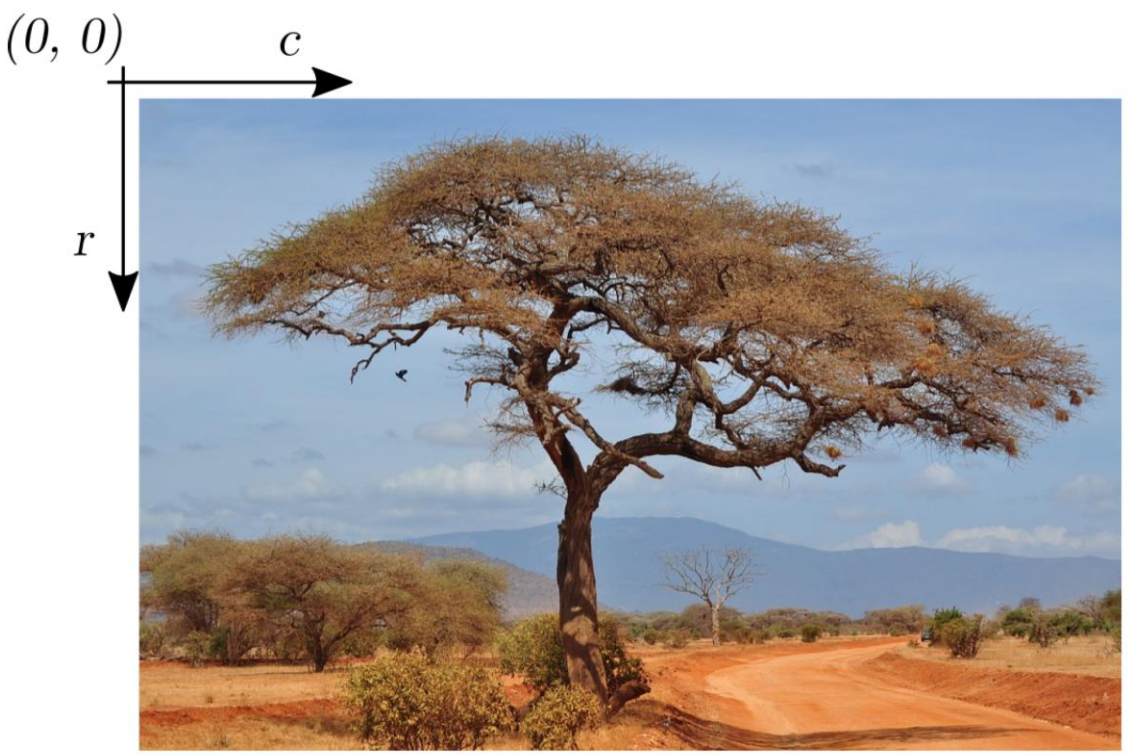
Example III/III

c_1: Tiger

c_2: Leopard

c_3: Lion

$$y = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ b_r \\ b_c \\ b_h \\ b_w \end{bmatrix}$$



$$\begin{bmatrix} 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ \varnothing \\ \varnothing \\ \varnothing \\ \varnothing \end{bmatrix}$$

Figure 19: Savannah. Image source: https://www.pixabay.com

Ø : We do not care!

Loss function

- Partition y as y = [c, b] with
  - c = [c_0, c_1, …, c_N] representing scores for all categories
  - b = [b_x, b_y, b_h, b_w] representing the bounding box
- Let c', b' be the ground truths.
- L_2 loss for the object's bounding box b

$$L_b(b, b') = \sum_{i \in \{x,y,h,w\}} (b_i - b'_i)^2$$

L1 or "smooth-L1" is also common

- Cross entropy loss for object categories c

$$L_c(c, c') = -\sum_{i=0}^{n} c'_i log c_i$$

- The total loss can be written as

$$L(y, y') = L_c + [c'_0 \neq 0] L_b$$

The terms can also be weighted

Compare bounding box only if there is an object

Pose estimation and landmark detection





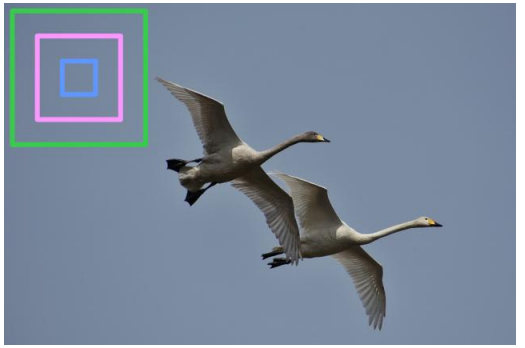
Sources:
Wu, Y., Ji, Q. Facial Landmark Detection: A Literature Survey. *Int J Comput Vis* **127**, 115–142 (2019). https://doi.org/10.1007/s11263-018-1097-z
Li, R., Li, Q., Yang, S. *et al.* An efficient and accurate 2D human pose estimation method using VTTransPose network. *Sci Rep* **14**, 7608 (2024). https://doi.org/10.1038/s41598-024-58175-8

Object detection (multiple objects)

- Unknown number of objects
- Make the network output *many bounding boxes* (and class-belongings)!

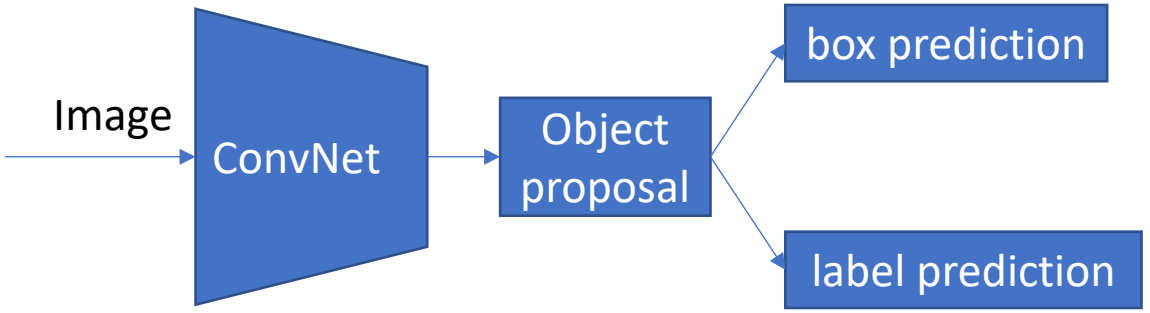
Naïve (Sliding window approach) for detecting multiple objects

- Slide (multiple-sized) windows across the image
- Apply an image classifier on every location
- OK for “cheap” classification methods
- Very slow for CNN classifiers

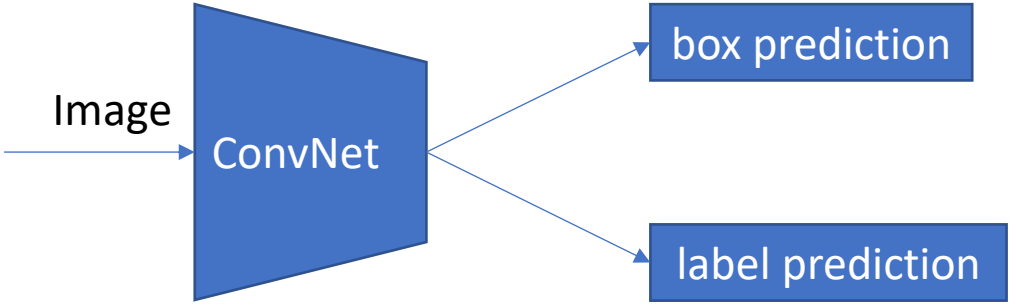


Object detection approaches

Two-stage detector



One-stage detector



Object detection approaches

Two-stage detectors

- R-CNN (2013 – 2014)
- SPP Net (2014 – 2015)
- Fast R-CNN (2015)
- Faster R-CNN (2015)
- R-FCN (2016)
- Feature Pyramid Network (2017)

One-stage detectors

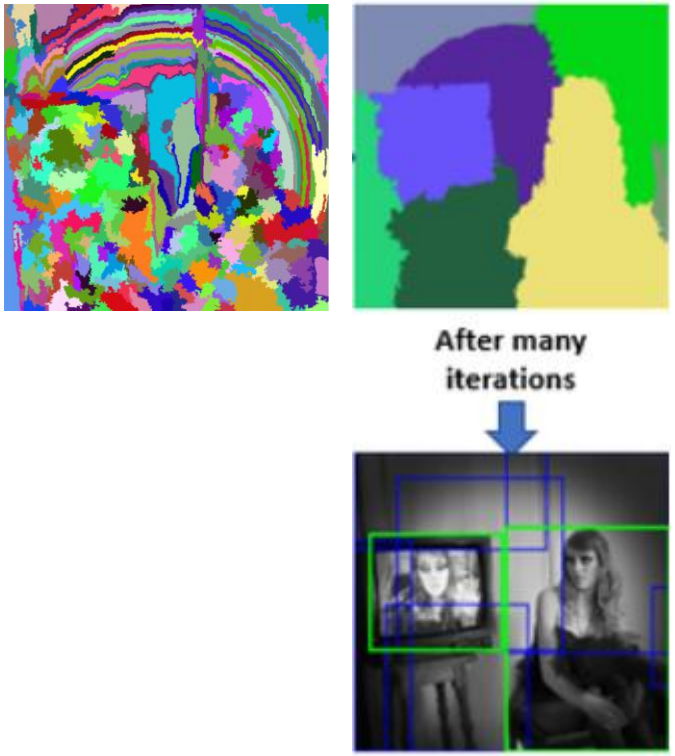
- YOLO (2015 – 2025)
  - Latest: YOLOv13 (2025)
- SSD (2016)
- RetinaNet (2017)
- CenterNet (2019)
- EfficientDet (2019 – 2020)
- Swin Transformer (2021)

Two stage detectors

- Stage 1
  - Generate multiple candidates (bounding boxes) (ROI) all over the image. It's typical to make use of "anchors".
    - Selective search, Region Proposal Network (RPN)
  - Throw away candidates without an object.
- Stage 2
  - Process each candidate independently
  - Assign a category (class) to each candidate
  - Adjust its bounding box.

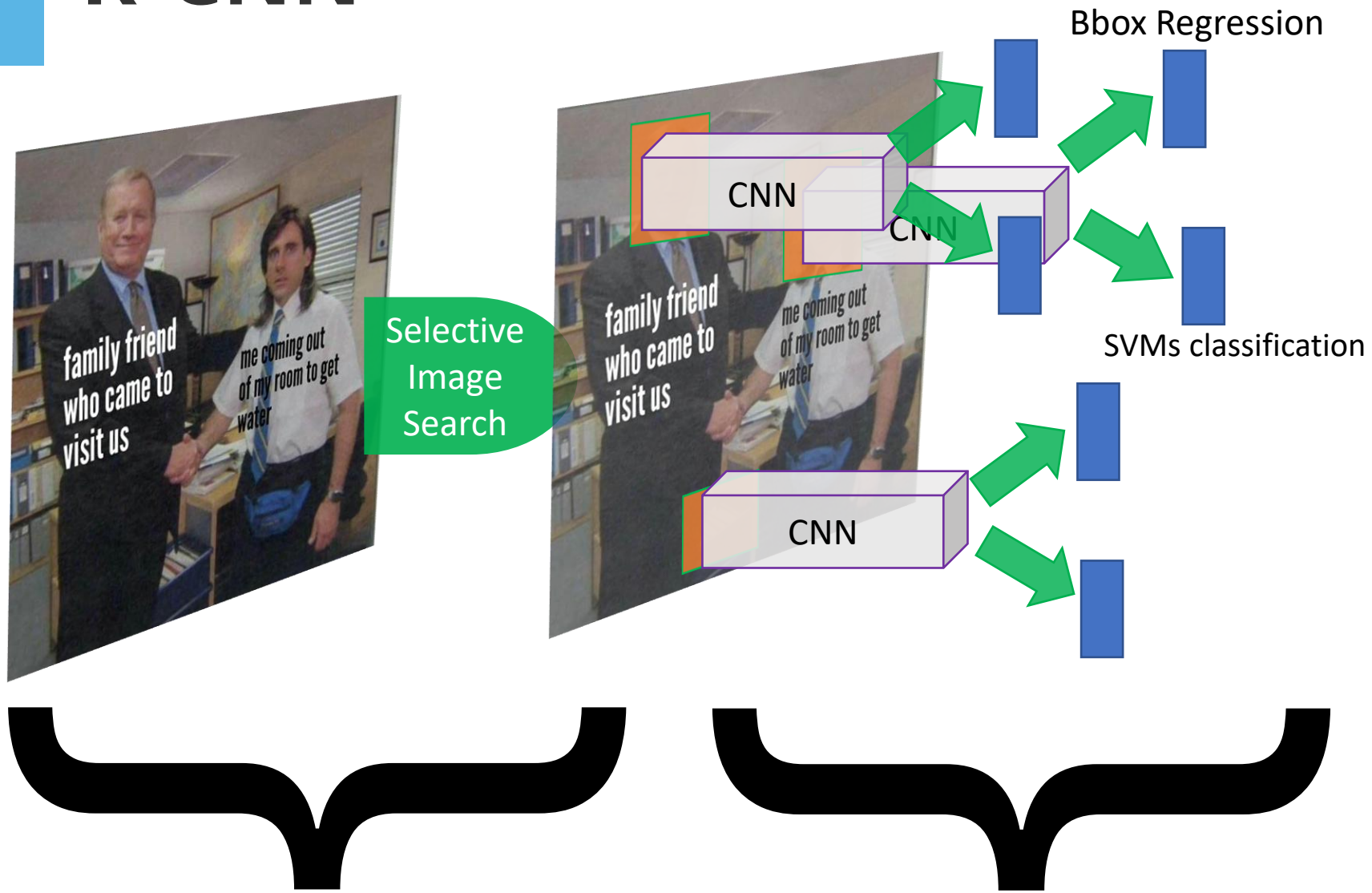
Region proposal

- Find specific candidates for windows
- Selective Search
  - No deep learning
- Region proposal networks



Source: Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, *104*(2)

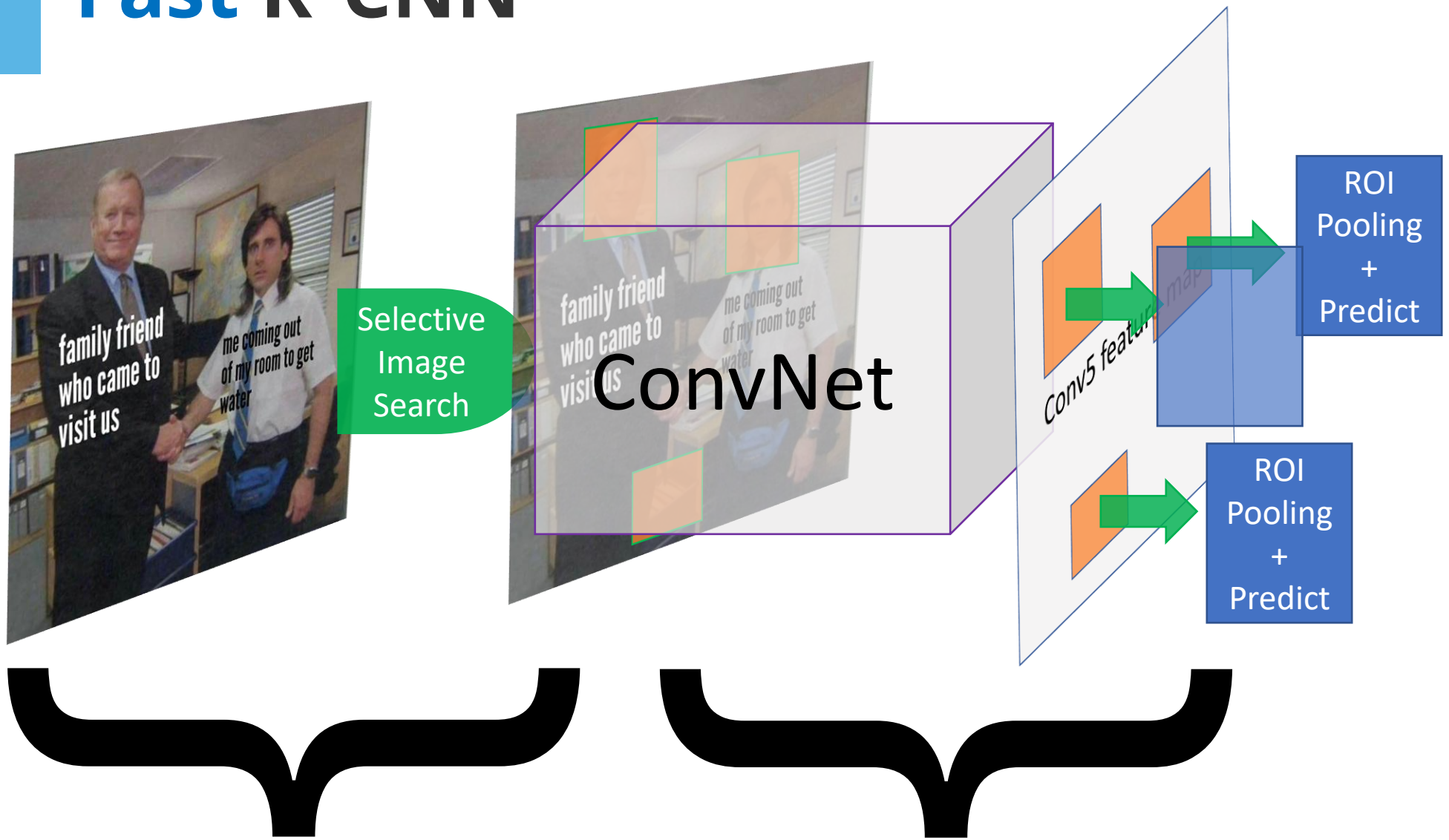
R-CNN



Stage 1 – Propose ROIs (Regions of Interest)

Stage 2 – Classify and adjust bbox

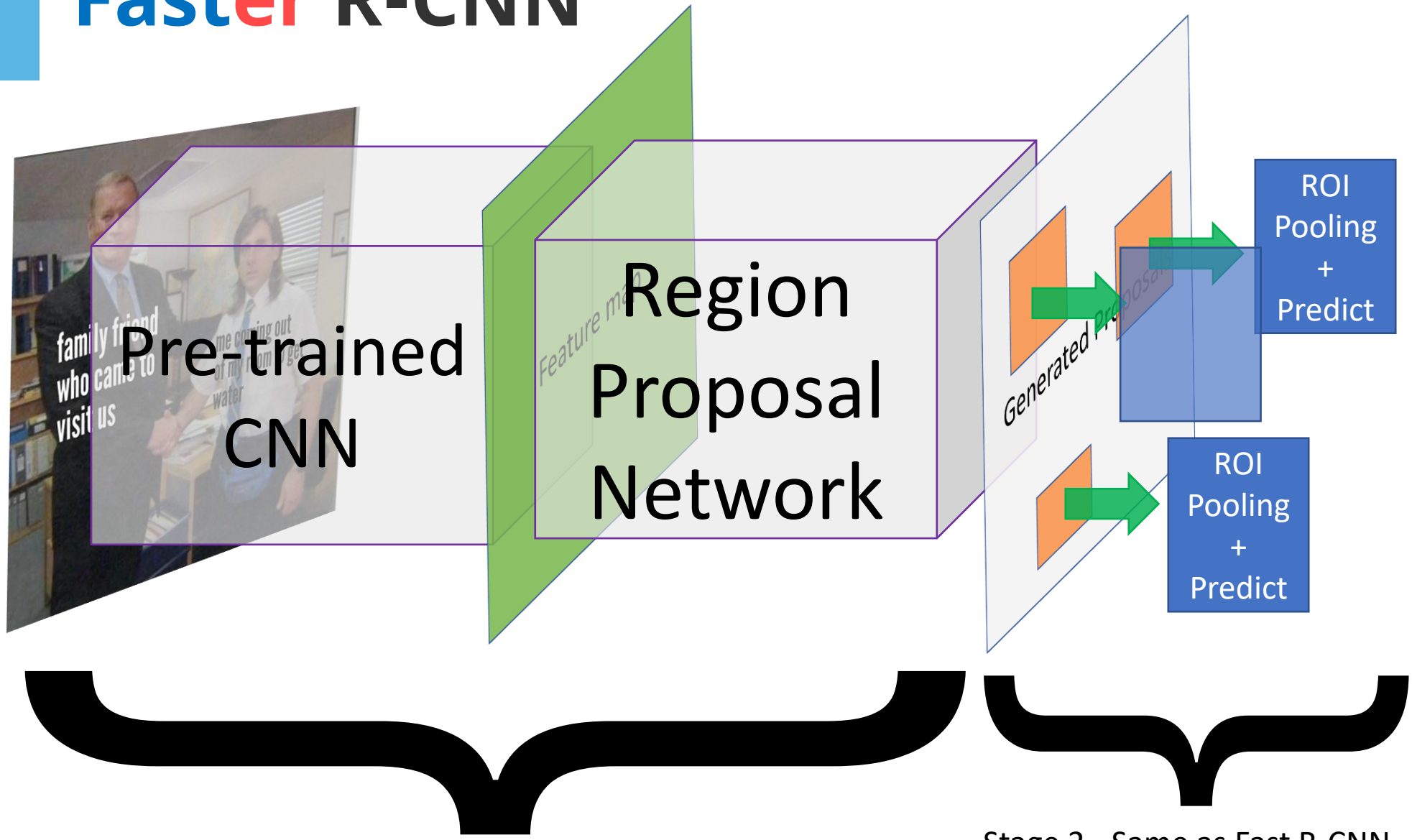
Fast R-CNN



Stage 1 – Propose ROIs (Regions of Interest)

Stage 2 – Classify and adjust bbox

Faster R-CNN



Stage 1 – Propose ROIs (Regions of Interest)

Stage 2 - Same as Fast R-CNN

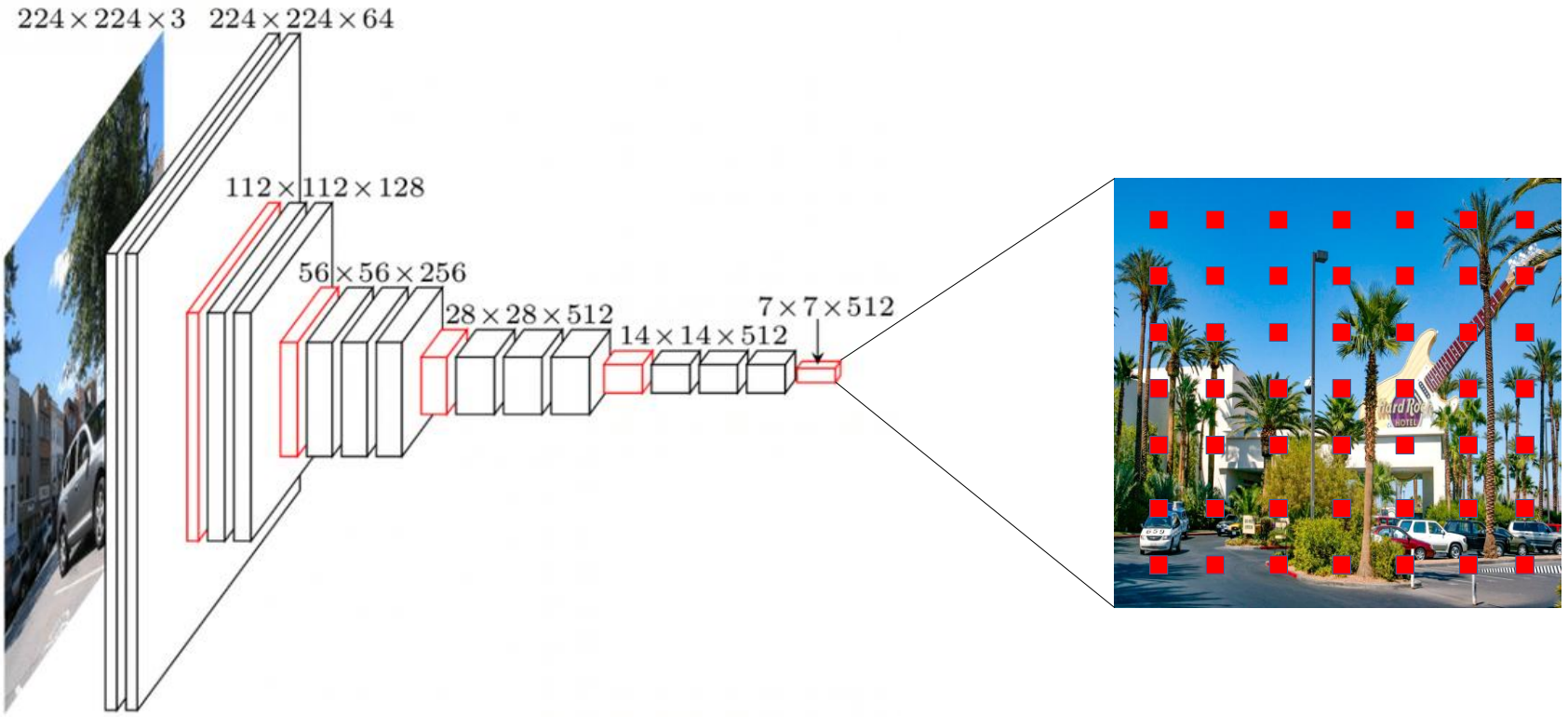
One stage detectors

- Pass (process) the image through a Neural Network.
- On the final feature map(s) of the network:
  - Slide a window
  - For each window location, predict an object class and a bounding box for each “anchor” (also called “default box” or “prior”), i.e., adjust the anchor.

One stage anchor-free detectors

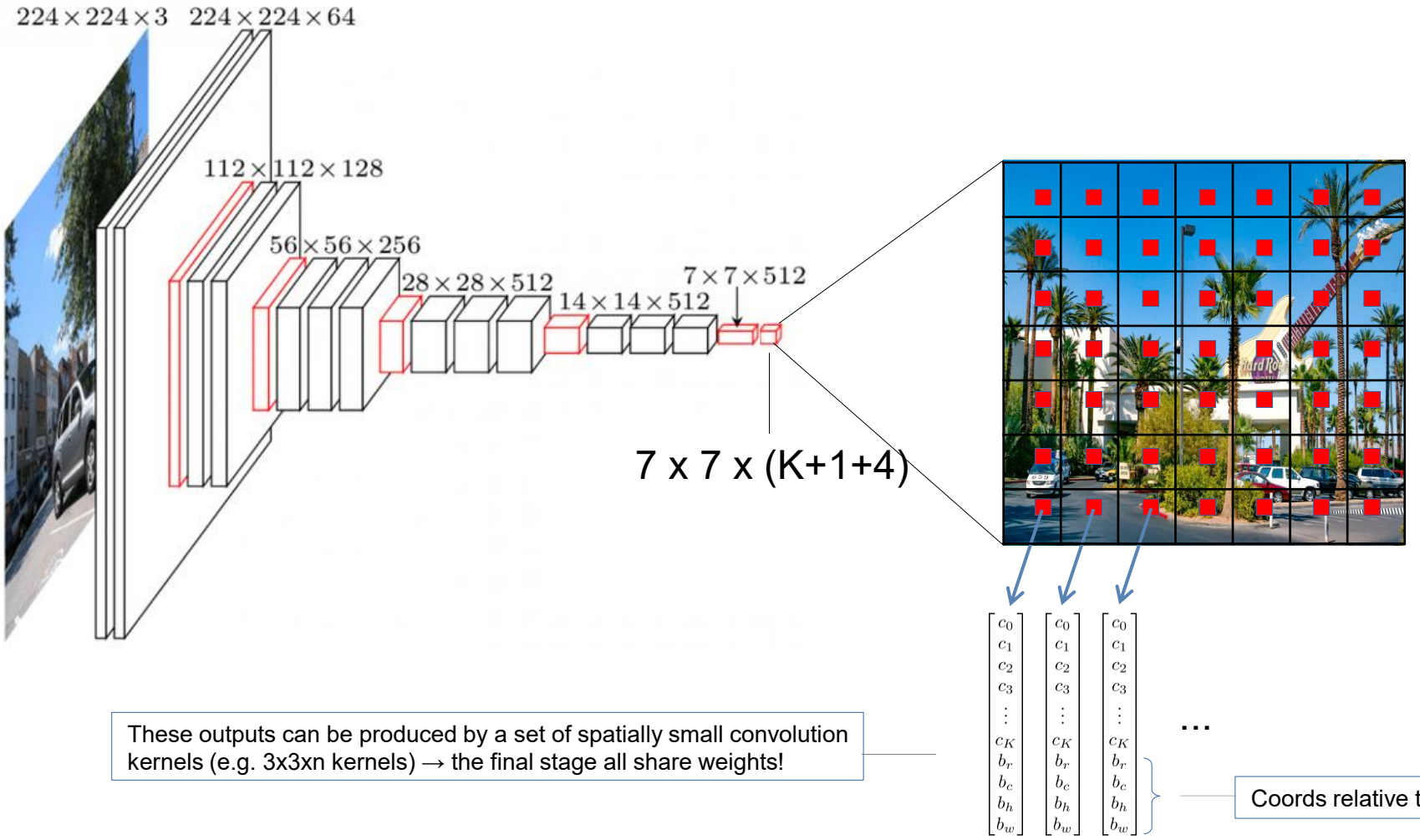
- Same as one-stage detectors, but instead of adjusting anchor, directly predict:
  - Top-left and bottom-right corners
  - And / Or the centre of the object

Generating candidates using anchors



Your network already produces features that are spatially distributed

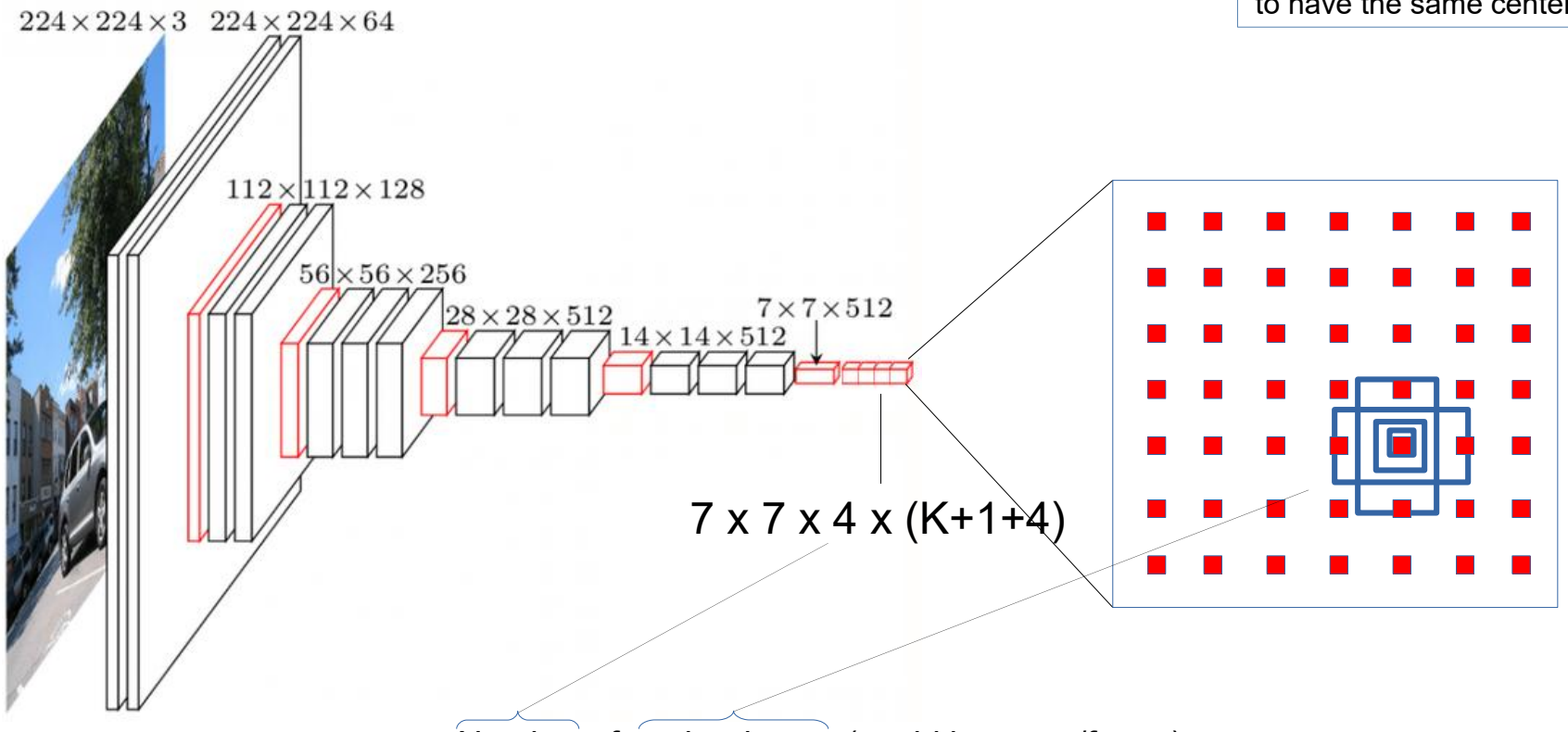
Generating candidates using anchors



These outputs can be produced by a set of spatially small convolution kernels (e.g. 3x3xn kernels) → the final stage all share weights!

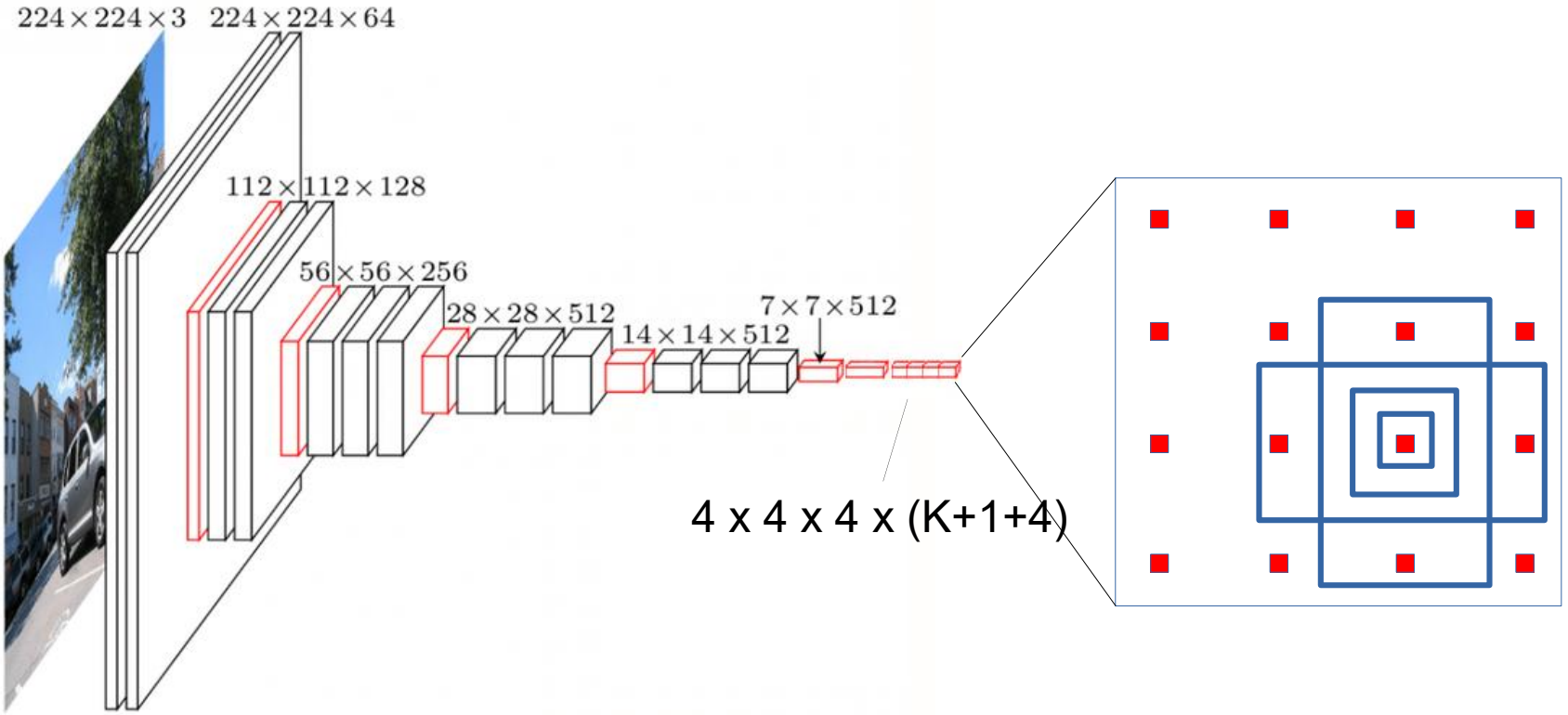
Generating candidates using anchors

Anchor boxes allow multiple objects to have the same center location



Sometimes anchor boxes go by the name *priors* or *default boxes*.

Generating candidates using anchors



Anchors on lower resolution feature maps can find bigger objects

Likely *fewer* bigger boxes needed

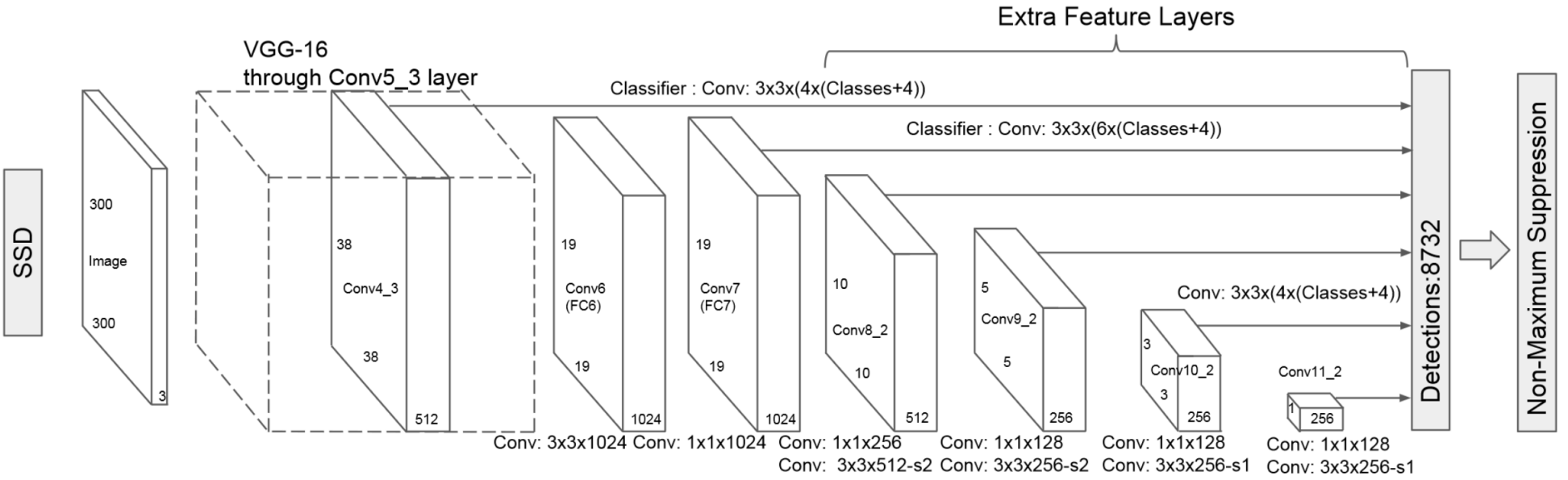
.. and, if still using 3x3 convolutions, the bigger boxes is predicted using a larger image context

Problem 1: Different object sizes

- Objects can be big or small across images or even within an image.
- Anchors of varying sizes help but anchors at a location are limited by that location's receptive field size.
- We'll look at two solutions (there are many):
  - Single Shot Multibox Detector (SSD)
  - Feature Pyramid Network (FPN)

Single-shot multibox detector (SSD)

- One stage detector



Liu, Wei, et al. "SSD: Single shot multibox detector", ECCV, 2016.

Even though this “simple” design works well in practice, we visit some common-seen refinements/improvements later.

Note that the YOLO algorithm, RetinaNet and SSD have strong similarities and are often grouped together (“single-stage detectors”).

Feature-pyramid networks (FPN)

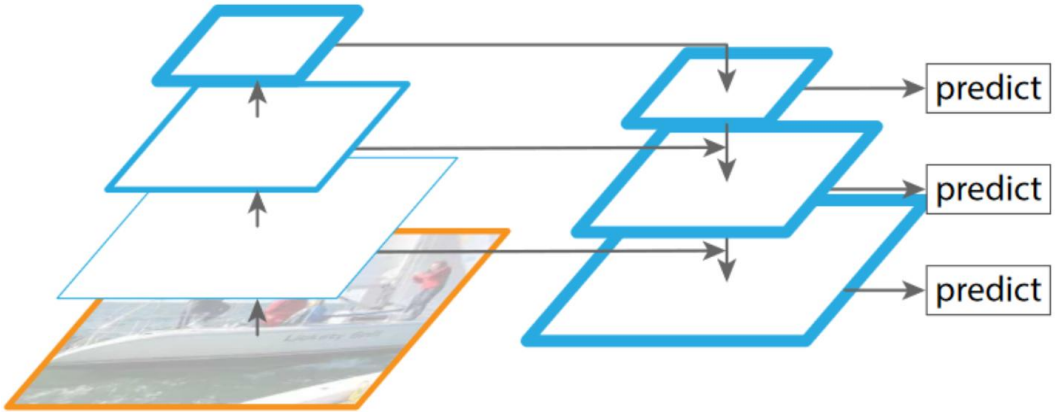
- Originally proposed as a two-stage detector.
  - Used as a backbone in one-stage detectors like EfficientDet.
- Deeper layers → semantically stronger features
- Let us try to propagate some of this “semantic strength” into the earlier layers before detecting our objects!

Note that there are many different ways of achieving this, however the ”FPN” described on the next slides is easily implemented, fast and does wonders in practice..

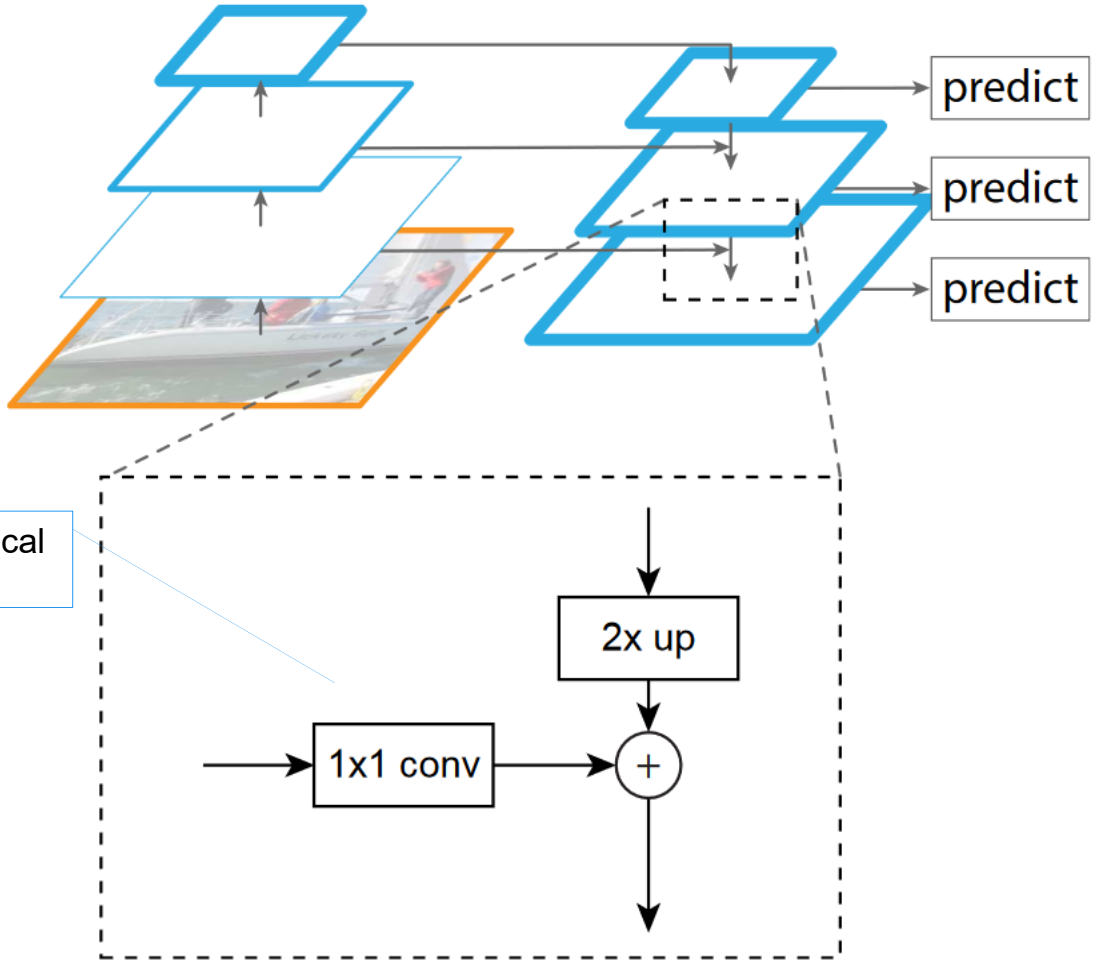
Context matters!



Feature-pyramid networks (FPN)



Feature-pyramid networks (FPN)



Ensures all layers have identical dimensions (e.g. d = 256)

Lin et al. "Feature Pyramid Networks for Object Detection", CVPR, 2017.

Problem 2: Multiple overlapping predictions



Problem 2: Multiple overlapping predictions

- An object in an image can have many overlapping boxes that cover it.
- How do we know which ones to throw away?
  - Use Non-Maximum Suppression (NMS).

Non-max suppression

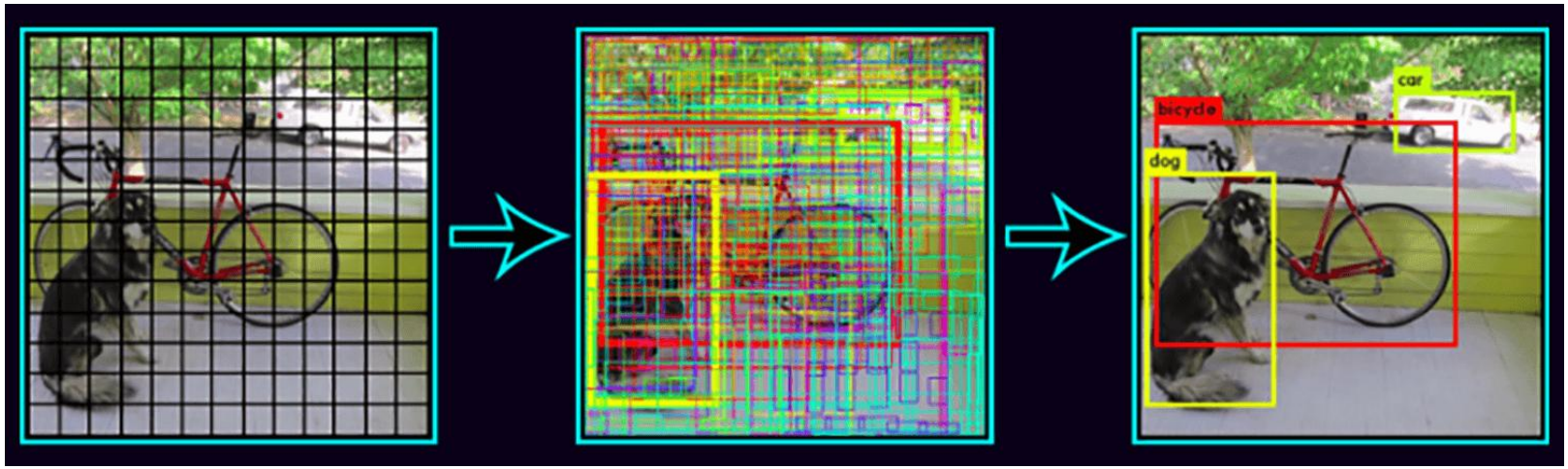
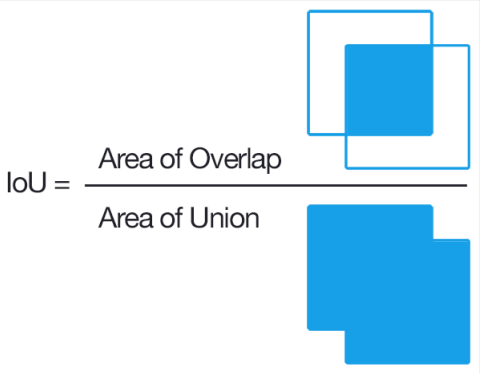
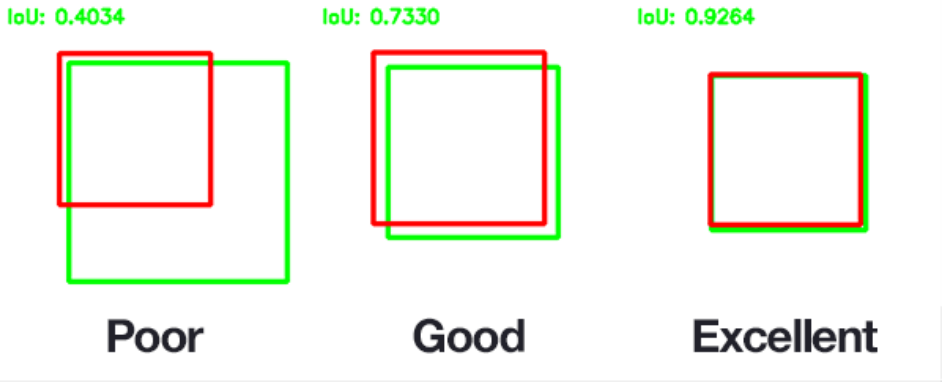


Illustration from the YOLO 2015 paper

Non-max suppression

Intersection over Union (IoU)





Illustrations: www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

Non-max suppression

- Important step in several object detection algorithms
- Remove all boxes having no c_1, c_2, .., c_K larger than, say 0.5

- For each class $i = 1, 2, \ldots, n$
  - Create a list of "unseen" regions U_i that contains all the regions in the image
  - Create an empty list of regions to keep K_i
  - While there are regions left in U_i
    - Find the most probable region $R_{\max}$
    - $R_{\max}$ can be the region with highest value of c_i (or some similar criterion)
    - Remove all regions that overlaps with $R_{\max}$ (e.g. with $iou > 0.5$), from U_i
    - Move $R_{\max}$ from U_i to K_i

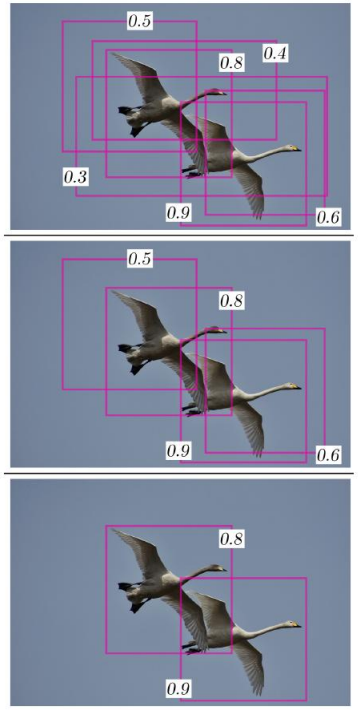


Figure 29: Top: Original. Middle: Too low c_0 removed. Bottom: $iou > 0.5$ removed. Image source: https://www.pixabay.com

Notes on how to train

How to match anchors with ground truth boxes?
Matching strategy used by SSD:

```
for every ground-truth box:
    match the ground-truth box with anchor-box having the biggest IoU

for every anchor-box:
    ious = IoU(anchor-box, ground_truth_boxes)
    max_iou = max(ious)
    if max_iou > threshold:
        i = argmax(ious)
        match the anchor-box with ground_truth_boxes[i]
```

That is, how to go from human annotated ground truth to “ideal” network outputs

Notes on how to train

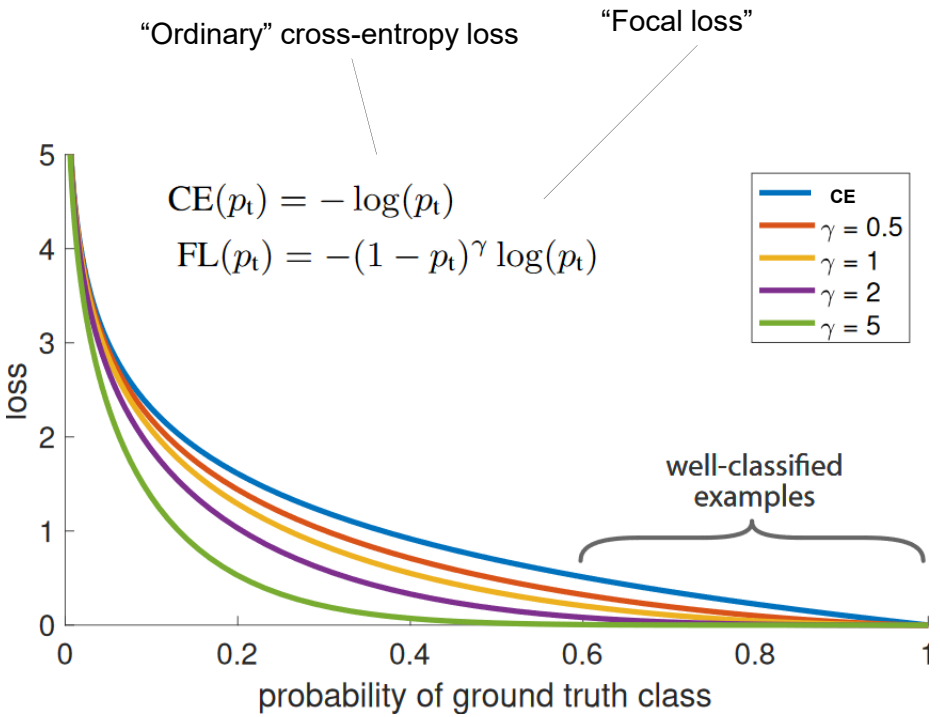
- Data augmentation (as always..)
  - Create more (and plausible so) data; cropping, resizing, mirroring, photometric distortions, …

Problem 3: Too many background predictions

- Often many more anchor-boxes without a match → *imbalanced dataset*
  - Hard Negative Mining
    - Select only *the most difficult* background patches (lowest background score) when computing loss.
    - See OHEM (Online Hard Example Mining) for instance.
  - Change loss function to downscale importance of highly certain background patches → "*focal loss*" (see next slide)
    - Mostly used by one-stage detectors

Focal loss

- "Focal loss .. focuses training on a sparse set of hard examples"
- Vigour of this focus controlled by a γ-parameter
- One can in addition add more weights to the non-background patches
  - Cf. the α factor often mentioned in conjunction with focal loss



Let's say we have many background patches which we are quite certain about being correctly labeled (p ≈ 0.7).

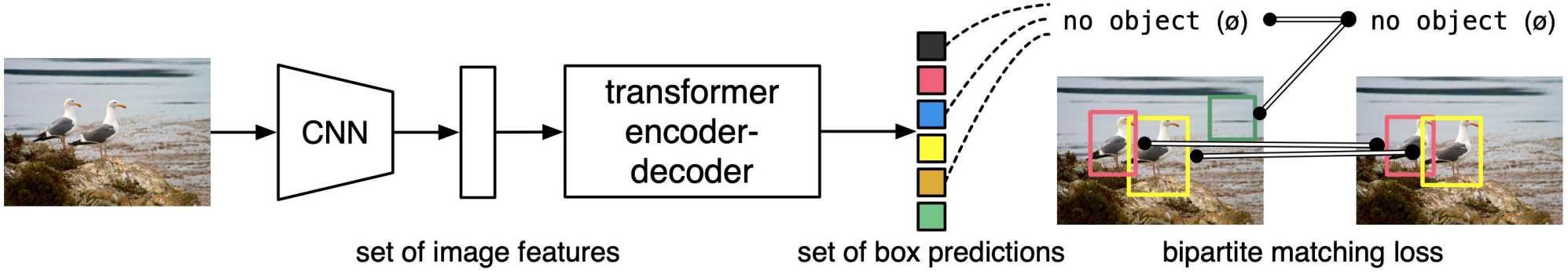
Using the CE-loss, we can substantially reduce this loss by going from "quite certain" to "very certain" (p ≈ 0.9), not so for the FL-loss.

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases}$$

Lin et al.. "Focal Loss for Dense Object Detection", ICCV, 2017.

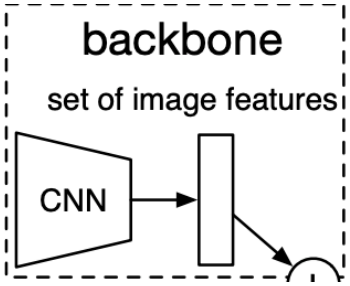
object DEtection with TRansformers: DETR

- Simple object detection pipeline: directly output a set of boxes from a Transformer.
- Anchor-free approach, i.e., no pre-defined anchors are used.
- No Non-Maximum Suppression (NMS) used. It is learned by the transformer automatically.
- Match predicted boxes to ground truth boxes with bipartite matching.



Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object detection with transformers: DETR

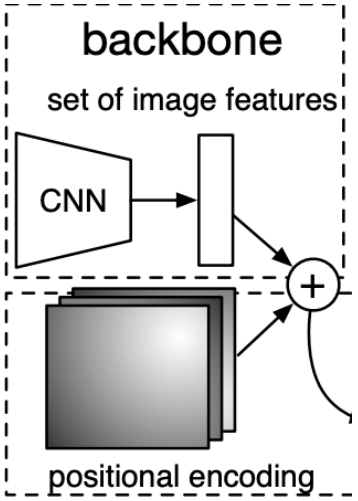


A backbone ConvNet (like ResNet50) is used to extract features from the input.

1x1 conv layer is used to reduce the number of channels.

Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object detection with transformers: DETR



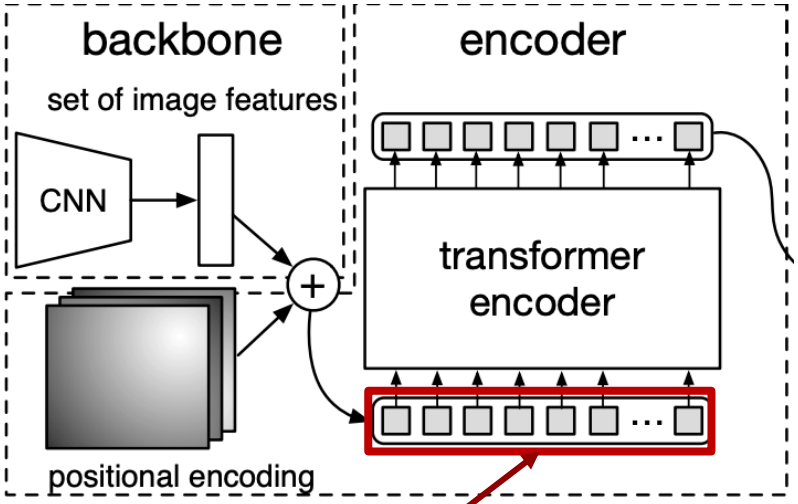
A backbone ConvNet (like ResNet50) is used to extract features from the input.

1x1 conv layer is used to reduce the number of channels.

Fixed positional encodings are added to every encoder block's input

Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object detection with transformers: DETR

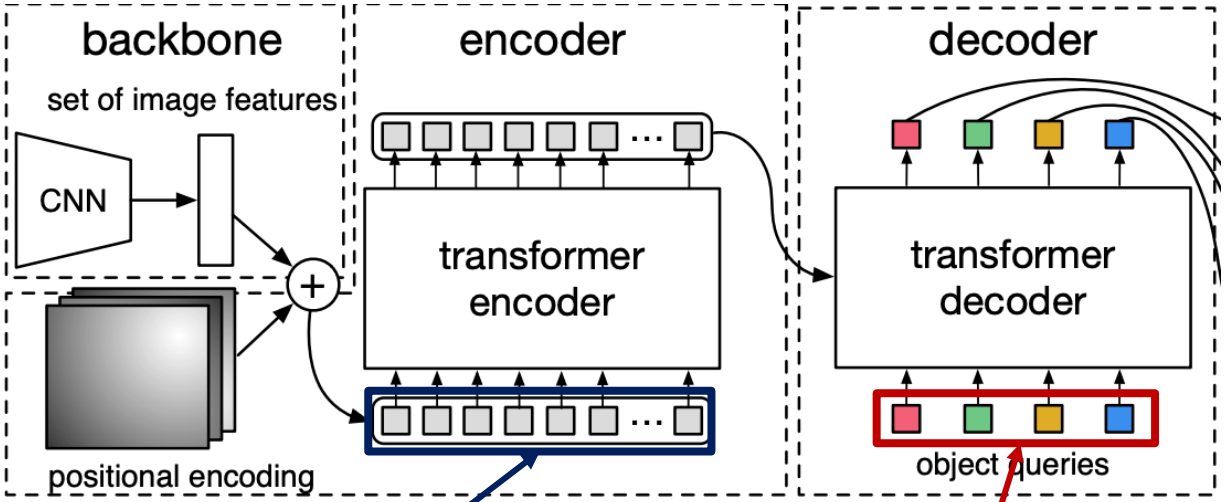


Standard transformer encoder is used to process the flattened image features.

HW x C
Flattened image features

Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object detection with transformers: DETR



HW x C
Flattened image features

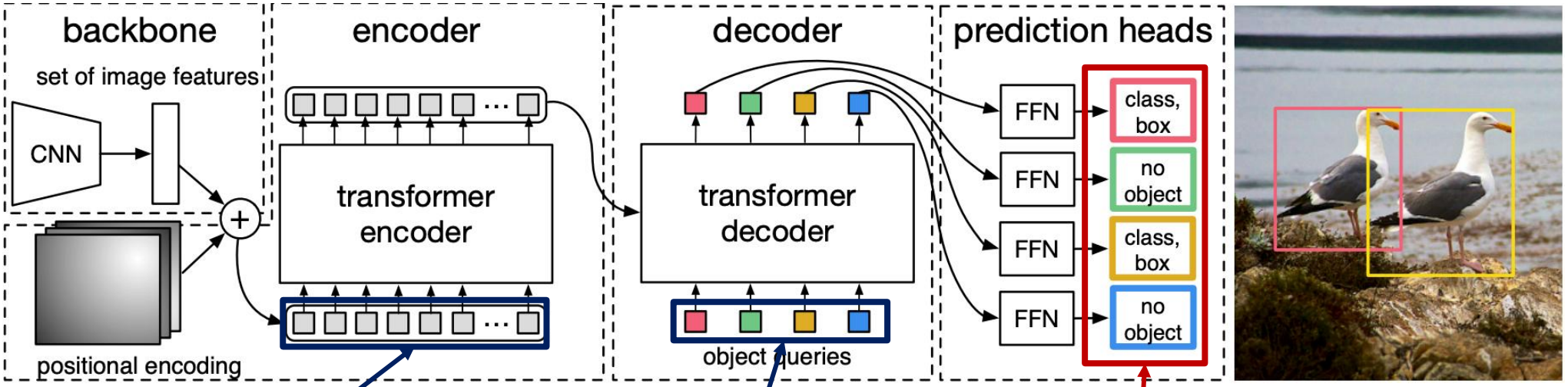
N x C
learnable object queries (randomly initialised)

A standard transformer decoder is used to generate outputs.

Input: N object queries, encoder's output

Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object detection with transformers: DETR



HW x C
Flattened image features

N x C
learnable object queries
(randomly initialised)

Each object query predicts one bounding box

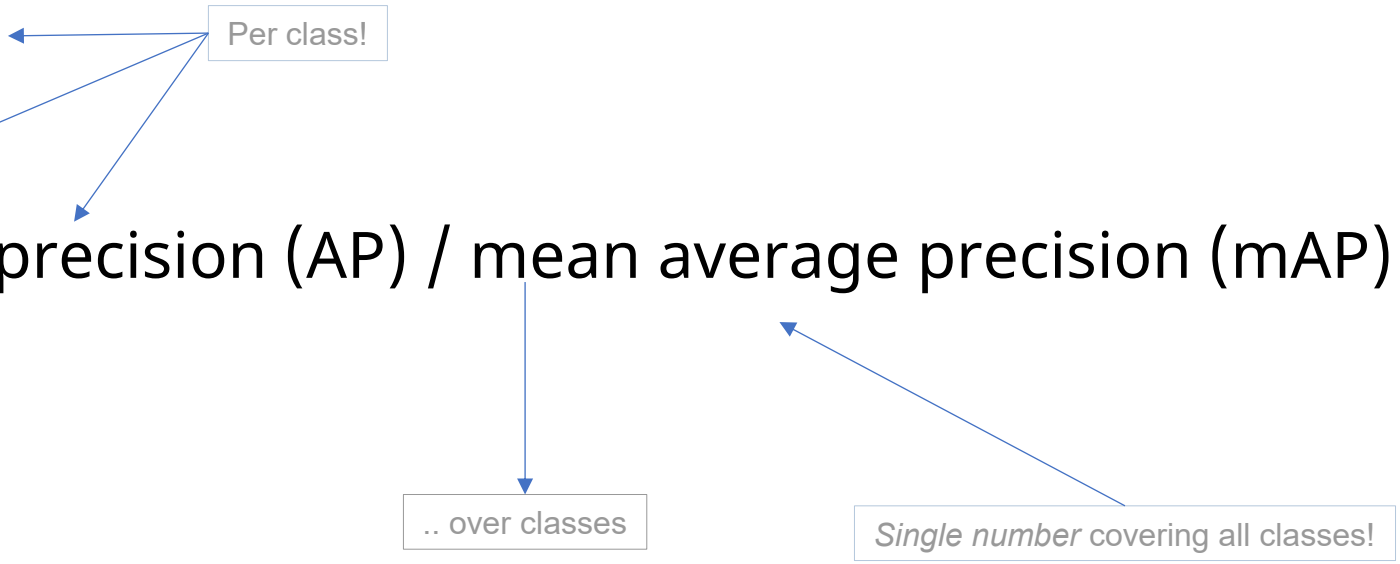
Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Near-infinite set of refinements..

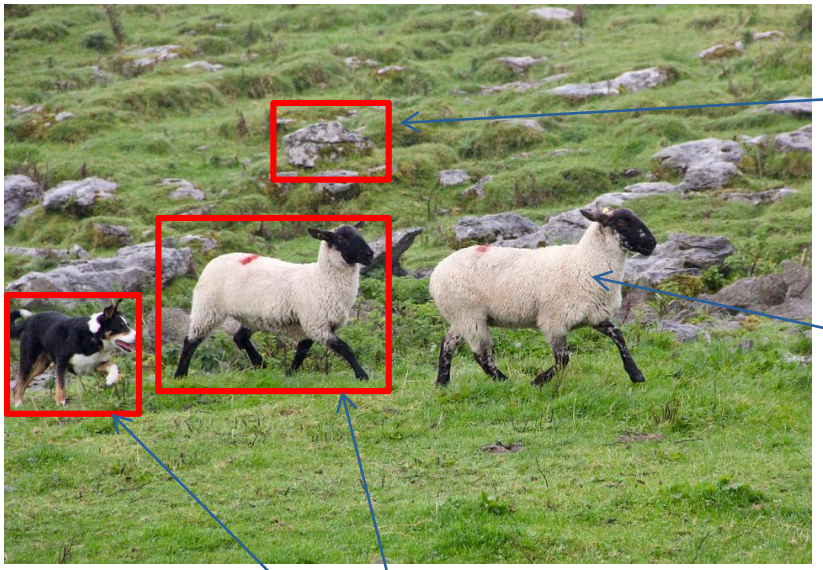
- “Backbones”
  - VGGs / ResNets / ConvNeXts / Transformers / ..
- Pretraining
  - What data and tasks are they trained on?
- Methods/implementations
  - SSD / RetinaNet / YOLO / YOLOv9 / ..
- Many versions and combinations of concepts (adaptive anchor-boxes, layer-merges, losses, learned NMS, …)
- Bag of Freebies – a bunch of things you can do that increase the training time but not the inference time.

Performance and evaluation metrics

- Precision
- Recall
- Average precision (AP) / mean average precision (mAP)



Assuming a ***single class***: “animal”



FP (false positive)

FN (false negative)

TP (true positives)

Note: Requires an ***IoU threshold***, e.g. 0.5 or 0.75.

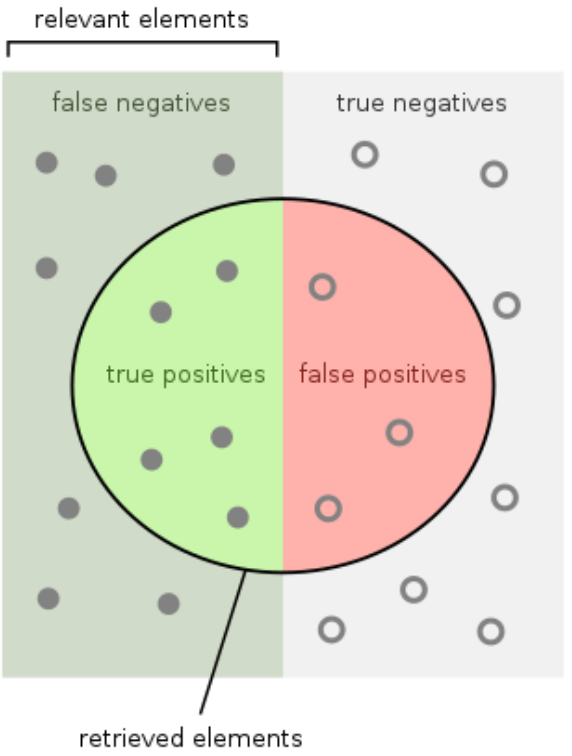
Remember: Precision/recall is calculated **per class**!

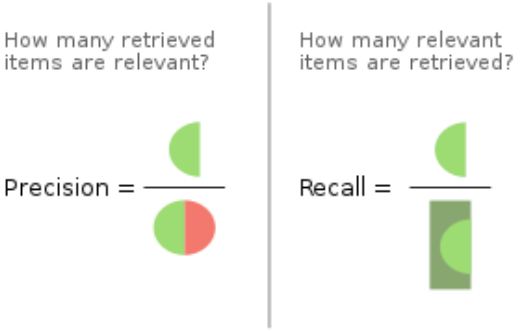
$$\text{Precision} = \frac{TP}{TP + FP}$$

(How many of our predictions are actually true/objects)

$$\text{Recall} = \frac{TP}{TP + FN}$$

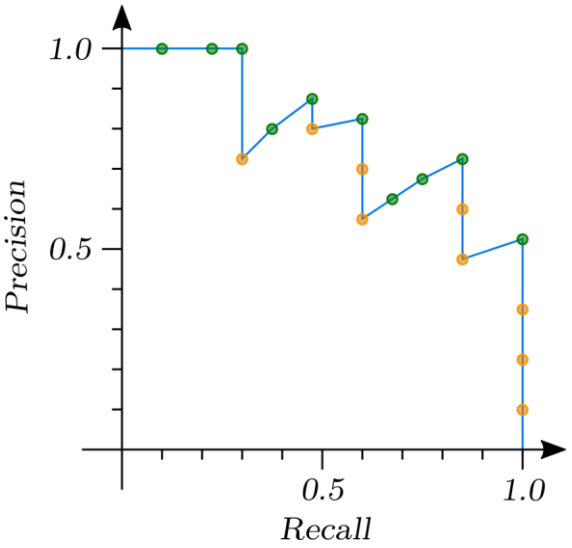
(How many of the objects of interest are found)





Precision-recall curves

- All detected objects have a class score, c_k, for each class k
- If we alter the score-threshold for what constitutes a predicted object of a given class, the precision and recall for this class will change
- Typically, lowering the threshold gives more detections, which increases the recall but lowers the precision
- By *changing this score-threshold,* we get a ***curve*** for each class

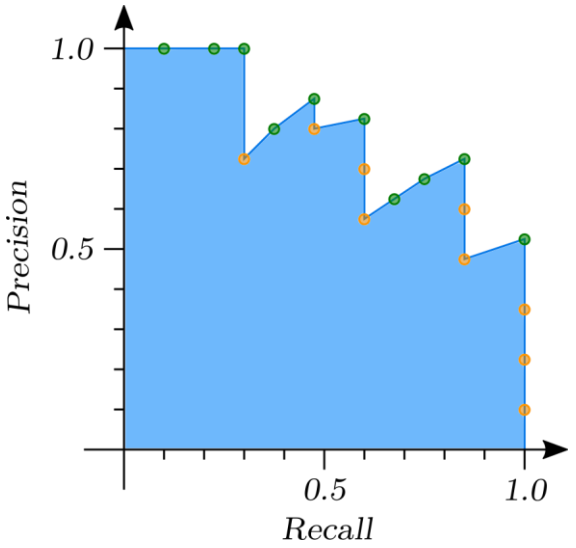


Remember that such curves depend on an ***IoU threshold.***

Average precision (AP)

- For a given class, the average precision is the **area** under its precision-recall curve (or some *approximation* of it)

- Note, we have an AP_k for each class k



Mean average precision (mAP)

The mean average precision (mAP) is simply the average AP over all the classes:

$$\text{mAP} = \frac{1}{K} \sum_{i=1}^{K} \text{AP}_i$$

One can also average over different IoU thresholds, e.g. as in the COCO and PASCAL VOC challenges.

In practice

- PyTorch/torchvision
- pytorch.org/hub/
- MMDetection, Detectron2
  - Actively developed toolboxes
  - A wealth of “backbones”, architectures and techniques
- Tensorflow object detection API

Other suggested resources

- University of Michigan's lectures on object detection:
  - https://www.youtube.com/watch?v=TB-fdISzpHQ
  - 1st half of https://www.youtube.com/watch?v=9AyMR4IhSWQ