AIMS | African Institute for Mathematical Sciences RWANDA

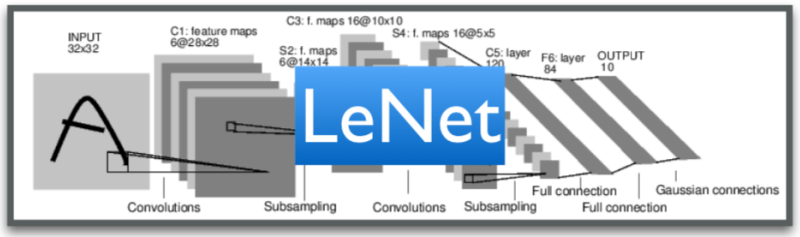
AMMI Review sessions

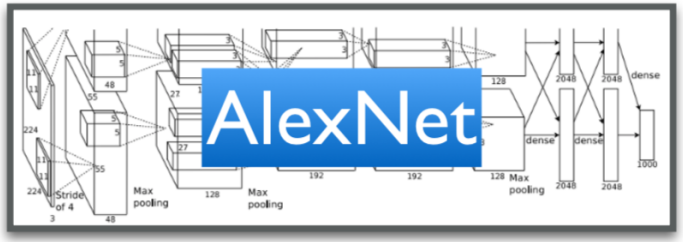
Deep Learning (6)

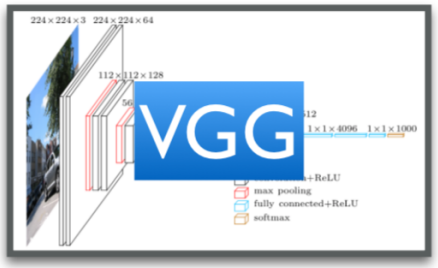
Convnets applications on visual recognition

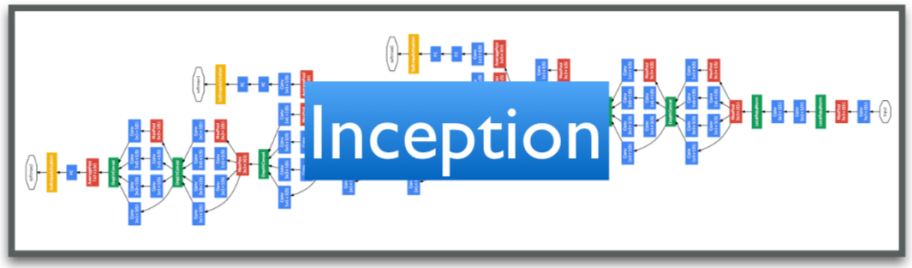
20 Feb 2020

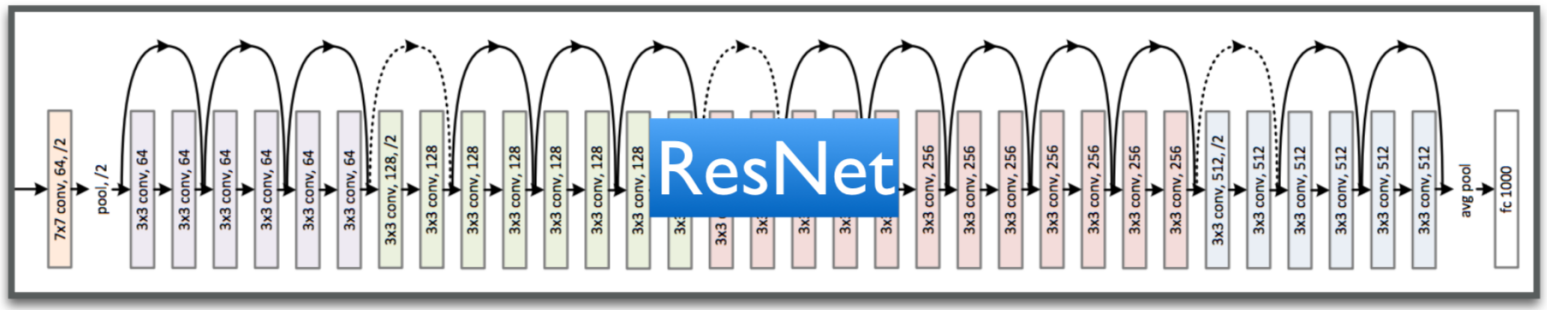
Popular CNN architectures









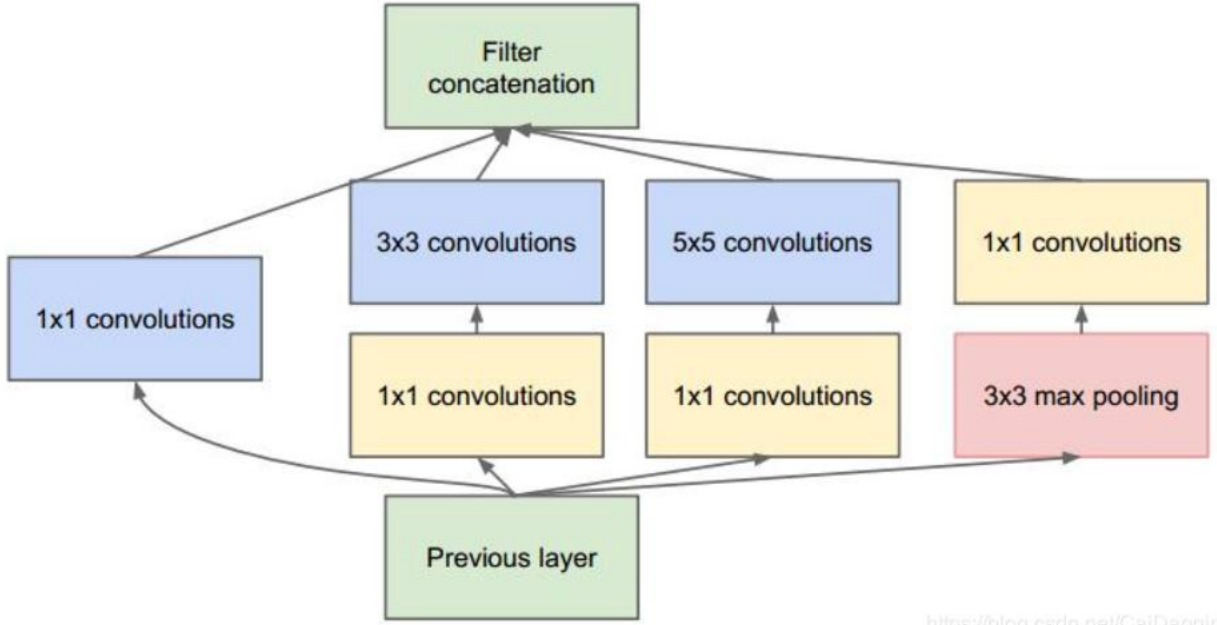


* Slide credits: Gao Huang

1x1 convolutions

- 1x1 convolutions are useful in the context of **non-linear** channel reduction. They allow us to preserve the width and height and reduce on the depth. They also require few parameters

- Inception module with dimensionality reduction



Global average pooling

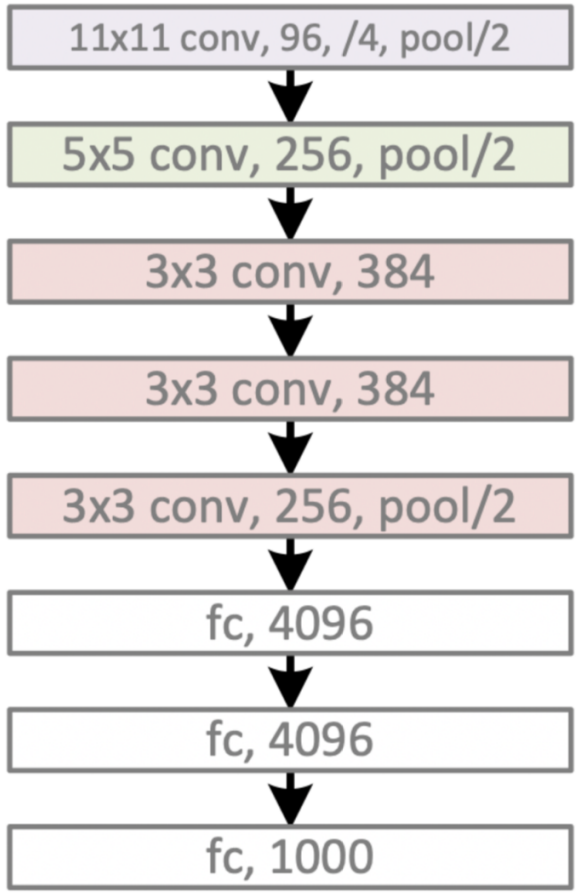
- Another method to flatten the volume is **Global Average Pooling**. It transforms a 3D volume to a 1D vector. It takes each 2D feature map and transforms them to singular values by taking the average of the feature map and concatenates them depth-wise

Popular CNN architectures

AlexNet

- LeNet-style model with three new components:
  - **Rectified linear units**
  - **Dropout** to reduce overfitting (currently less popular)
  - **Data augmentation** (still very important)

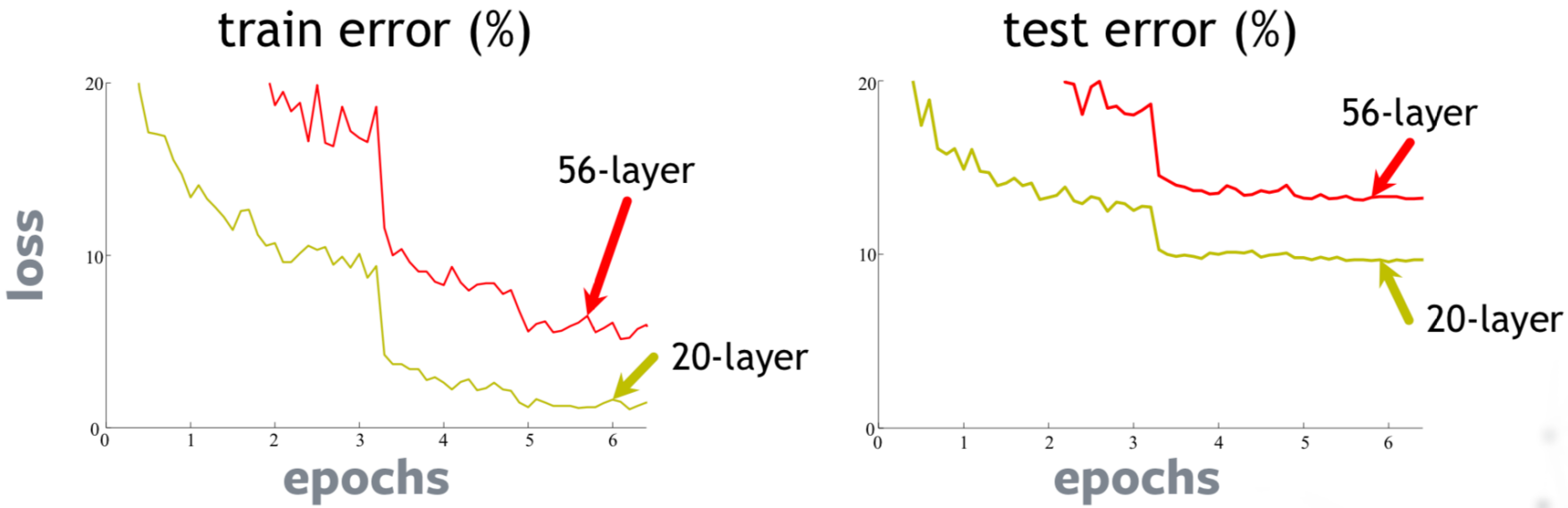
- AlexNet played a key role in popularizing convolutional networks



Popular CNN architectures

Stacking layers

- Simple experiment with stacking many 3x3 convolutional blocks:

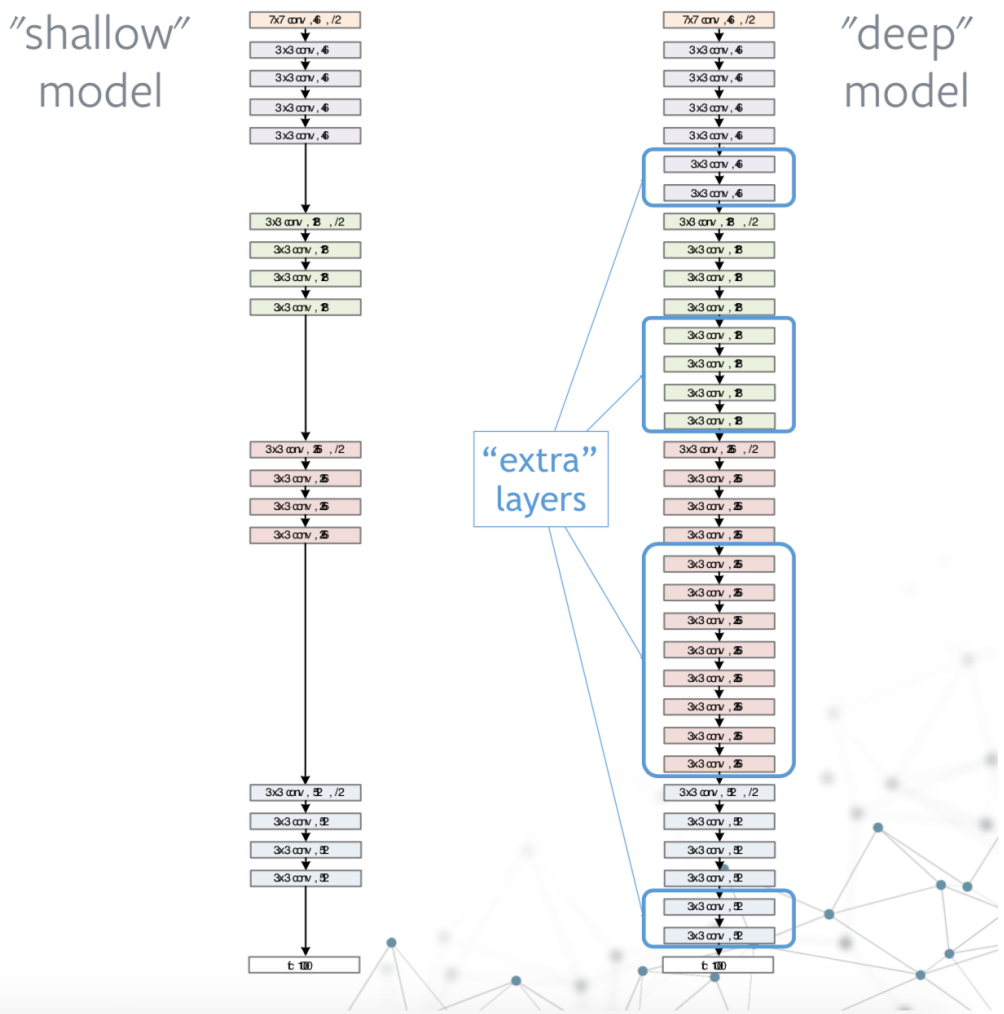


Popular CNN architectures

Stacking layers

- Deeper models should not have higher training error

- Solution by construction:
  - Train shallow model
  - Add additional layers set to identity
  - Train the deeper model further

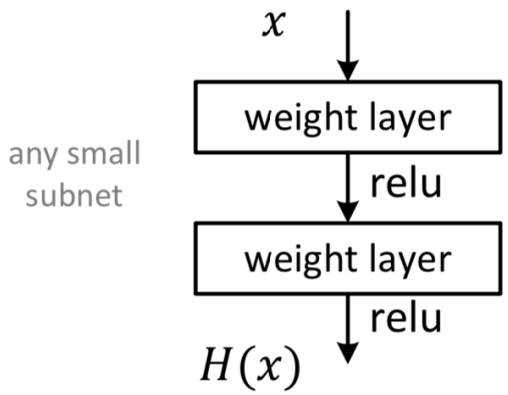


* Figure credit: Kaiming He

Popular CNN architectures

Residual connections

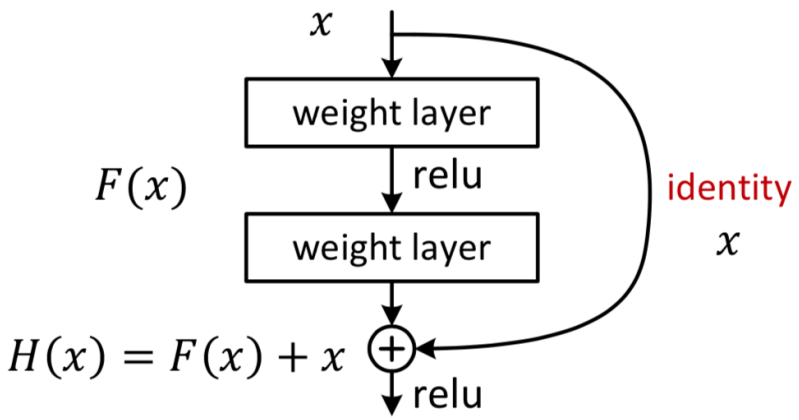
- Block architecture that achieves similar goal
- Take any "regular" network block...



Popular CNN architectures

Residual connections

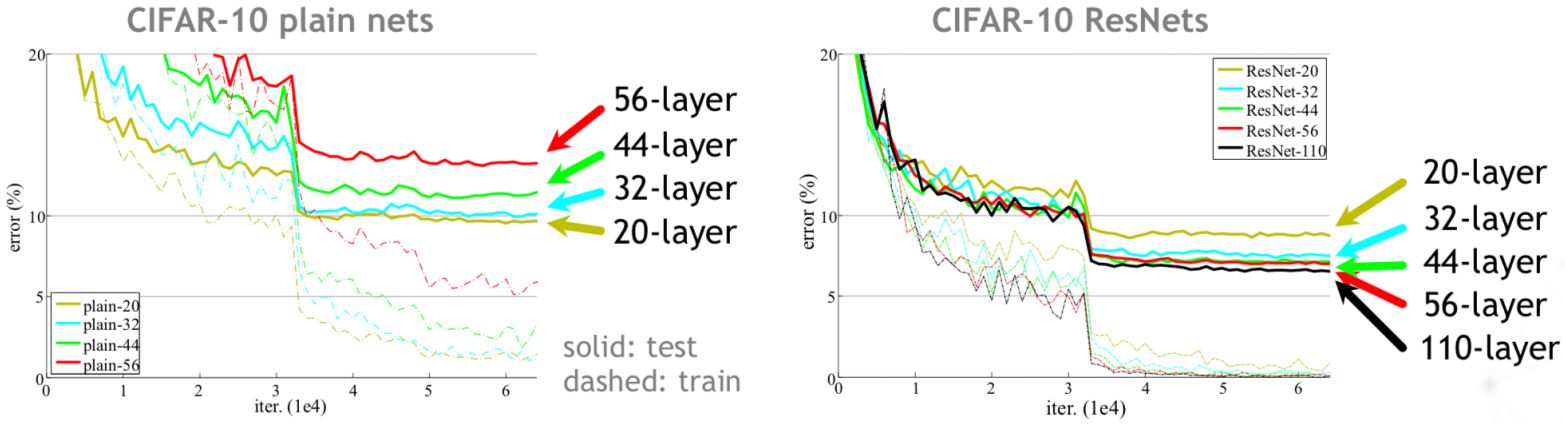
- Block architecture that achieves similar goal
- Take any "regular" network block... and make it **residual**:



Popular CNN architectures

Residual connections

- Deeper models can now be trained without problems:



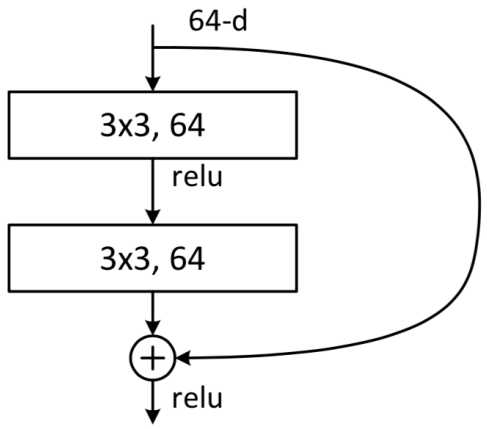
- Overfitting may still happen, but at least training error does not go up

* Figure credit: Kaiming He

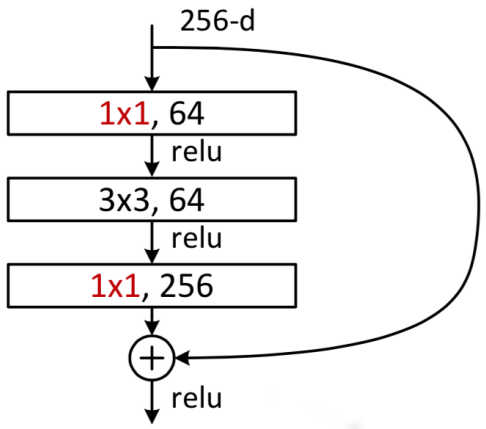
Popular CNN architectures

Residual networks

- In practice, residual networks (ResNet) use the following block design:



simple design



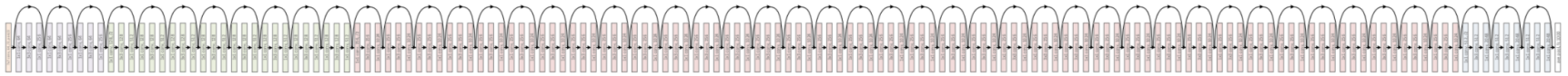
bottleneck design

* Figure credit: Kaiming He

Popular CNN architectures

Residual networks

- Full model has pooling layers for downsampling
- All pooling layers do max-pooling but the last one does average-pooling

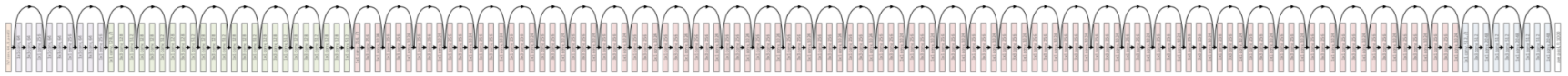


- Every time the image size halves, the number of channels is doubled

Popular CNN architectures

Residual networks

- Full model has pooling layers for downsampling
- All pooling layers do max-pooling but the last one does average-pooling



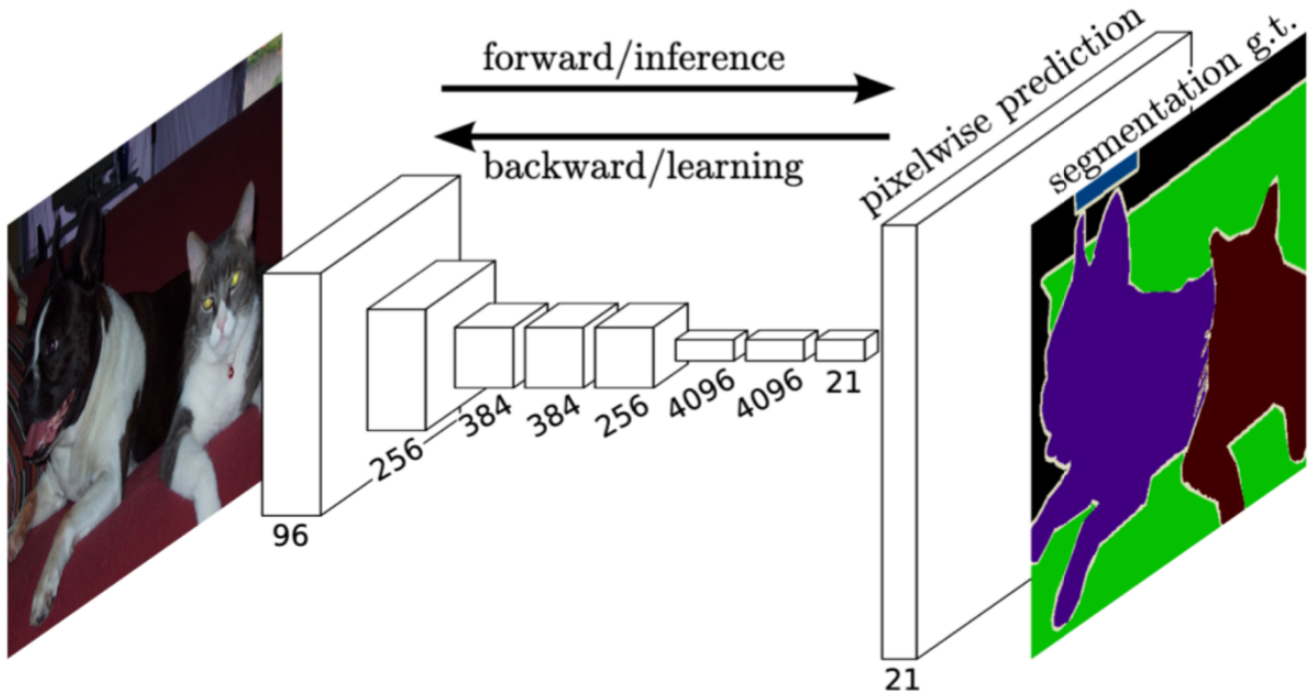
- Every time the image size halves, the number of channels is doubled

Visual Recognition tasks

Semantic Segmentation

- Given an input image I, mark each pixel in I with a class label
- Unlike image classification, predictions are cast for every pixel
- Metrics:
  - Mean Accuracy
  - Mean Intersection over Union

FCN



- Input image I of size H x W x 3
- Output logits of size H x W x C, where C is the number of object categories in the dataset.

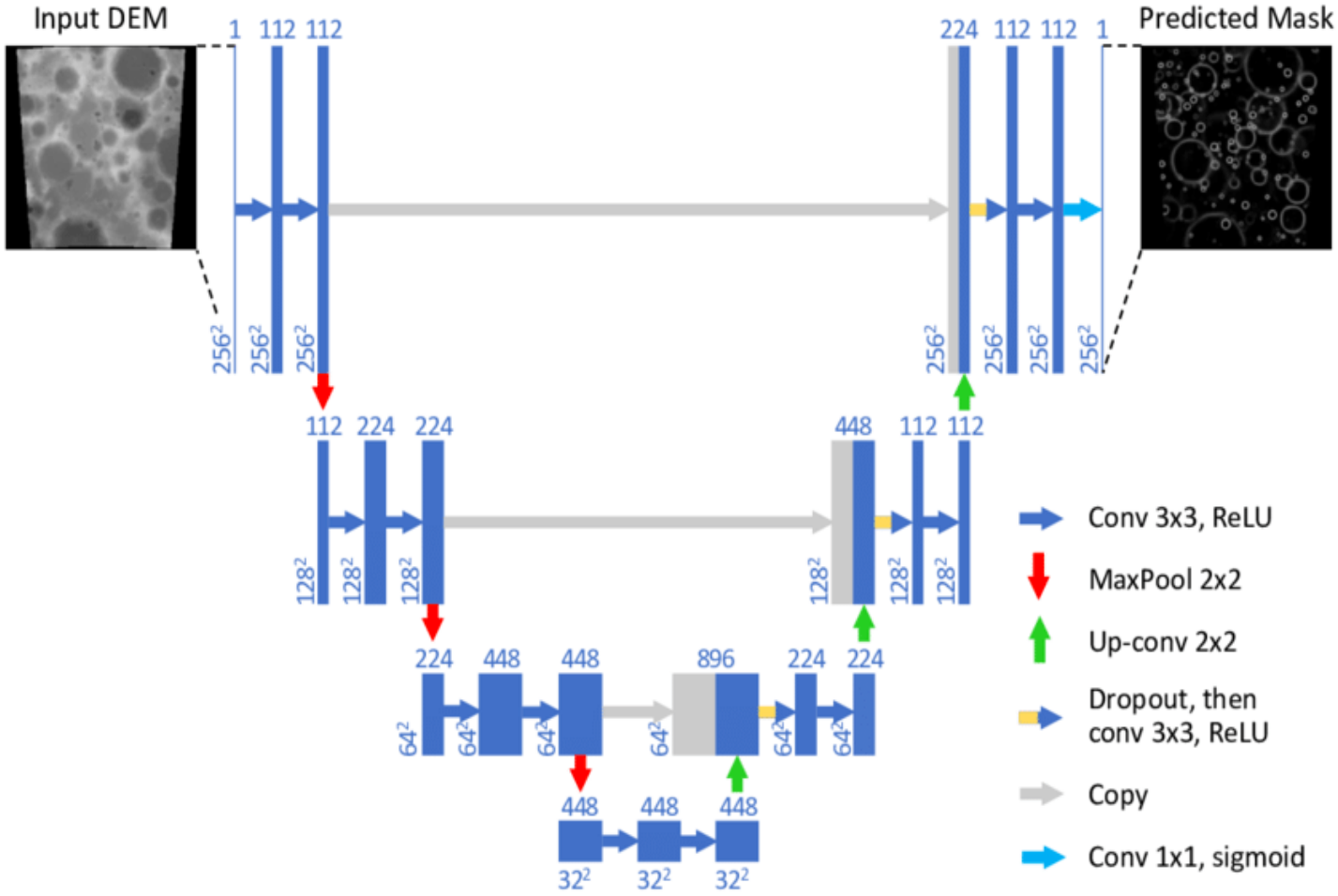
Long et al, FCN for Semantic Segmentation, CVPR 2015

FCN: Losses

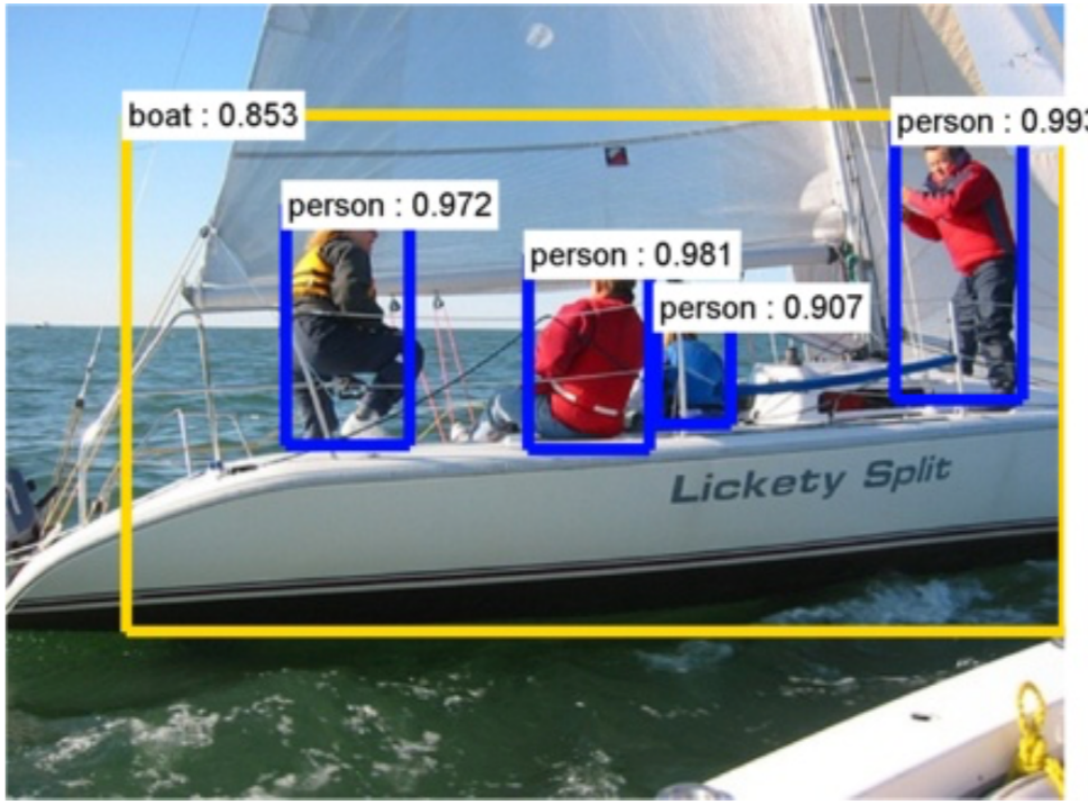
- Assume J $\in \mathbb{R}^{HxWxC}$ are the logits predicted by FCN & $J_{gt} \in [0, C-1]^{HxW}$ is the ground truth
- Binary classification vs. Softmax

U-Net

- An encoder decoder network with skip connections
- Capture the global context and localize low frequency components
- Reducing the need to large number of training examples because it applies elastic deformation to the training samples which makes the network invariant to such deformations without being explicitly present in the training data.



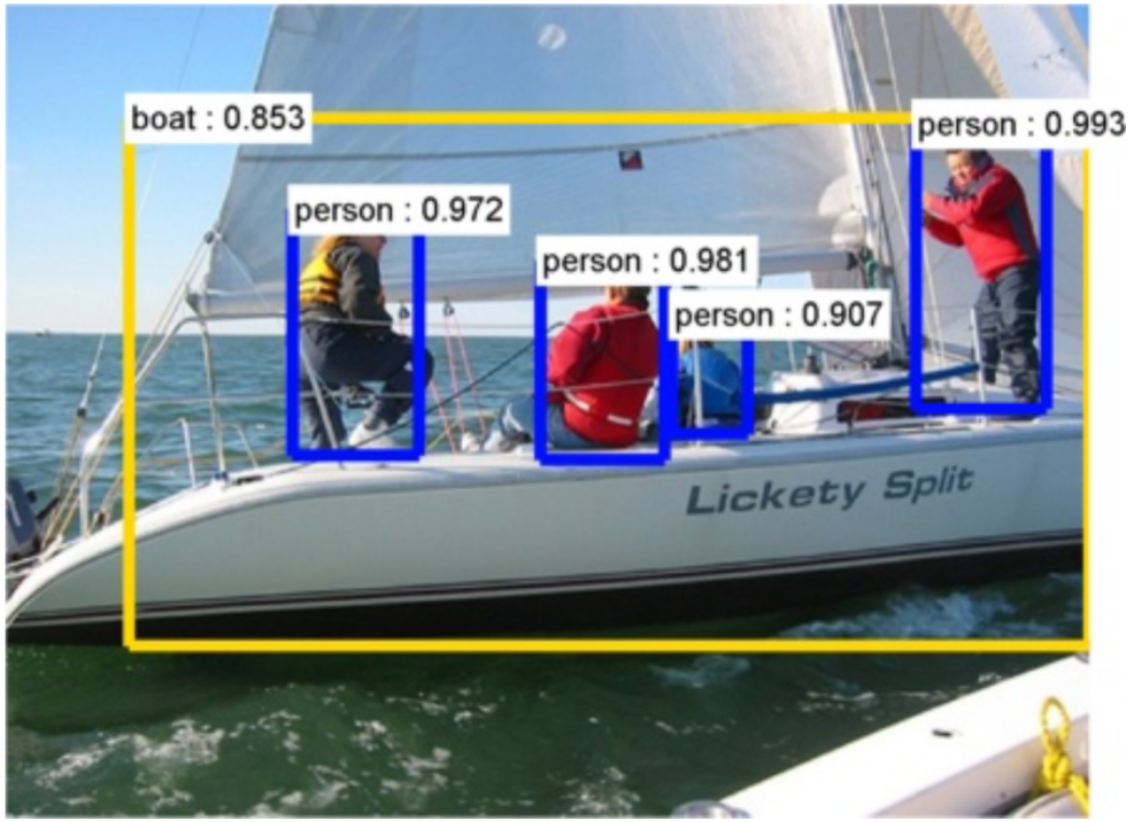
Object Detection



What

Where

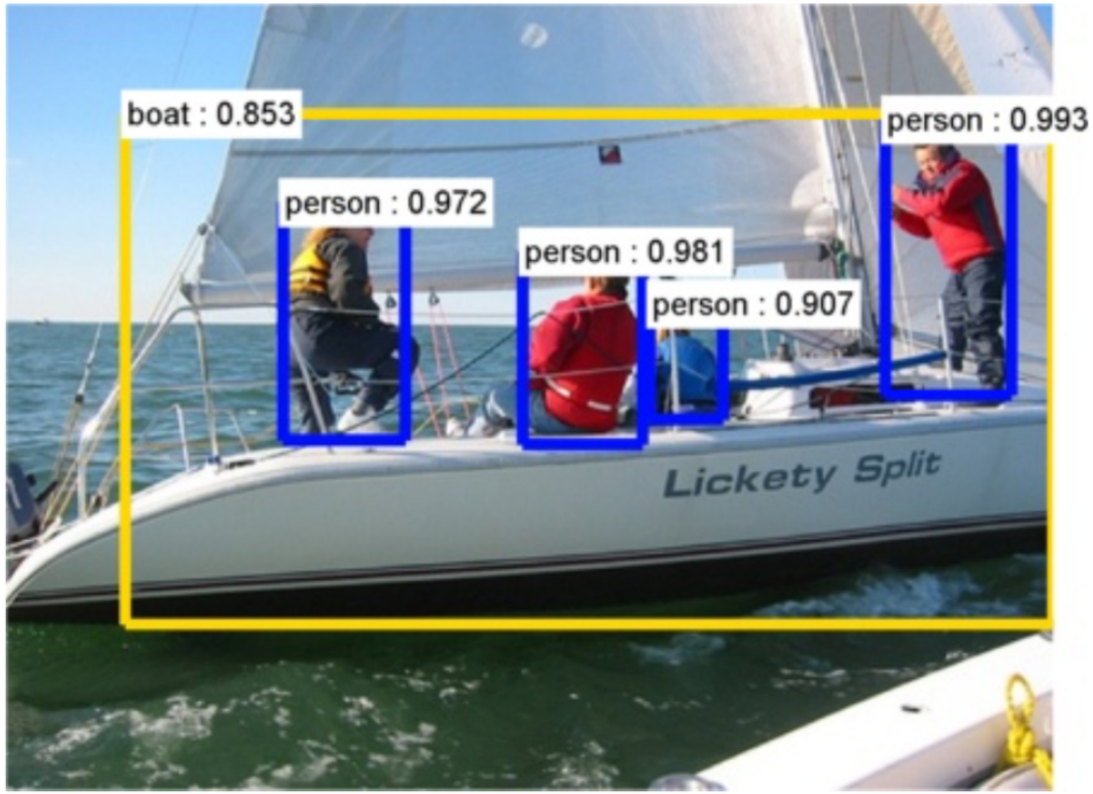
Object Detection



Object Detection

Image Classification

Object Detection



Task

- Assume C object classes
- Localize and classify all objects in an image

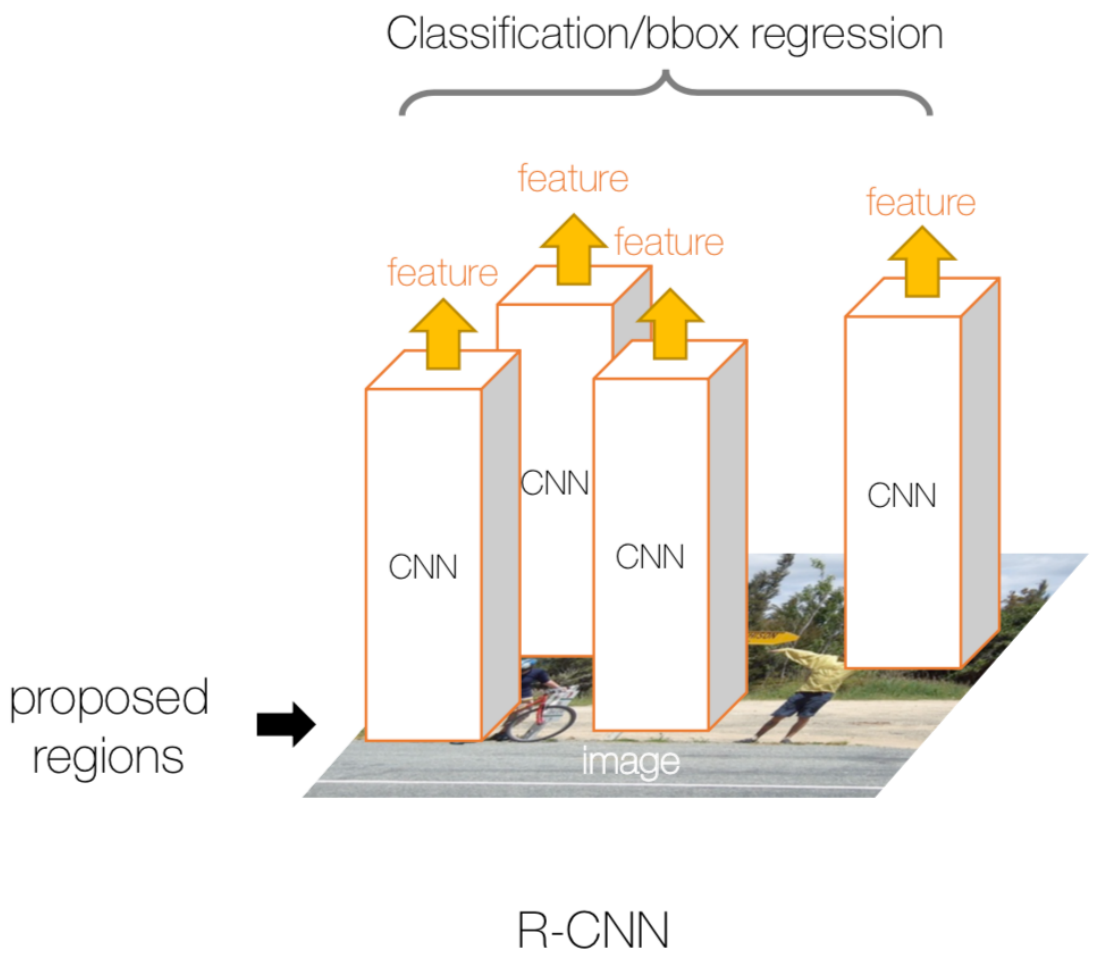
True Positives

- A detection b is a true positive if ov(b, gt) > 0.5 and class(b) = class(gt)

False Positives

- Mislocalized and misclassified detections
- Duplicate detections are penalized

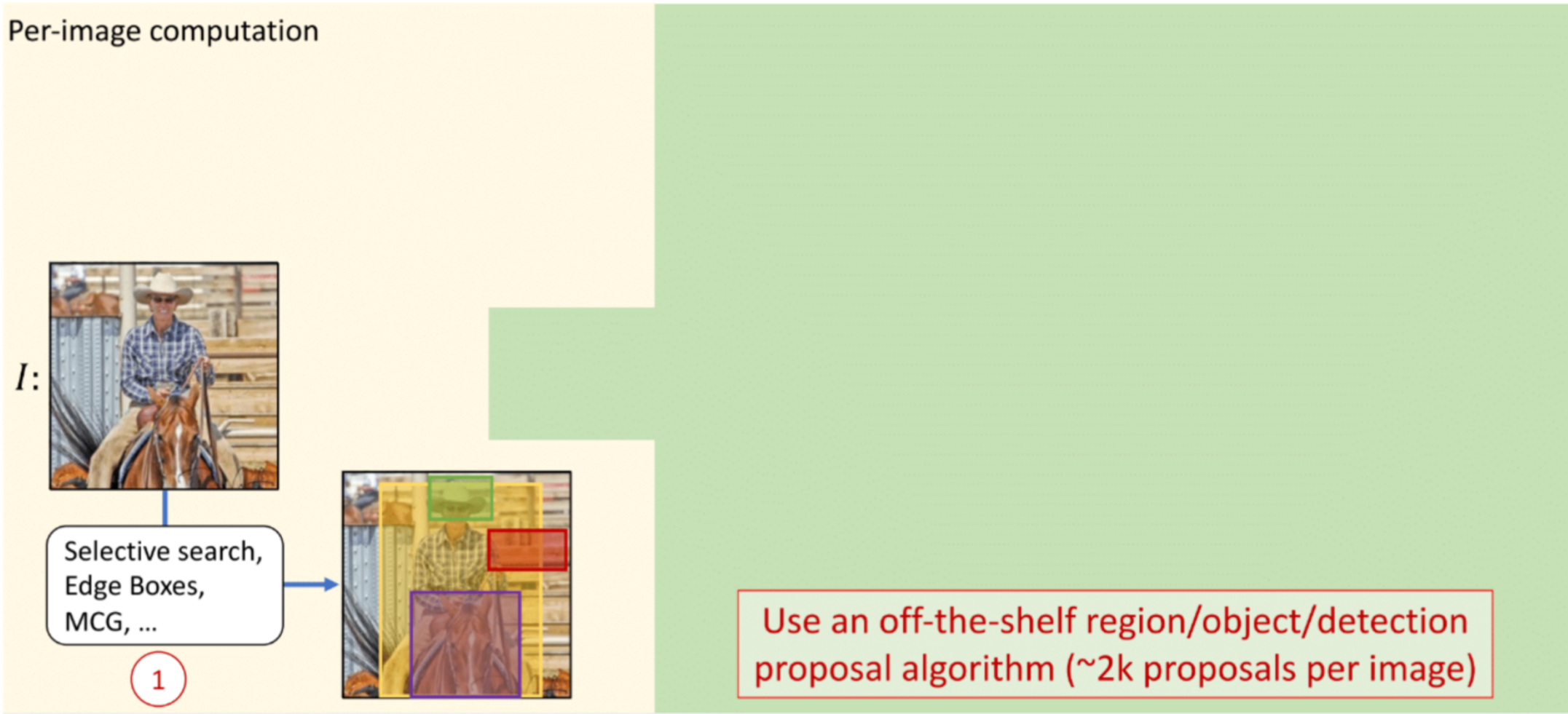
Object Detection Frameworks: R-CNN



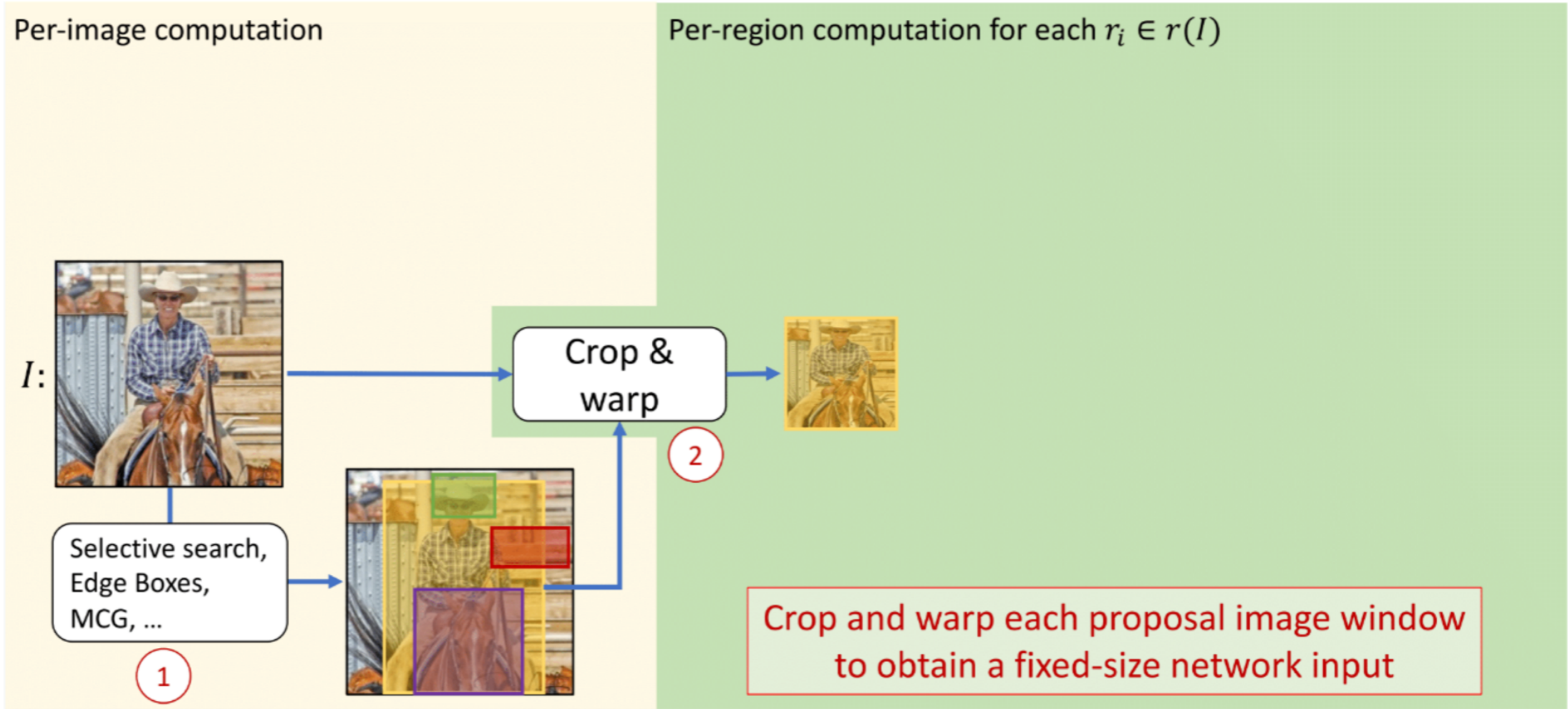
R-CNN

Girshick et al, CVPR 2014

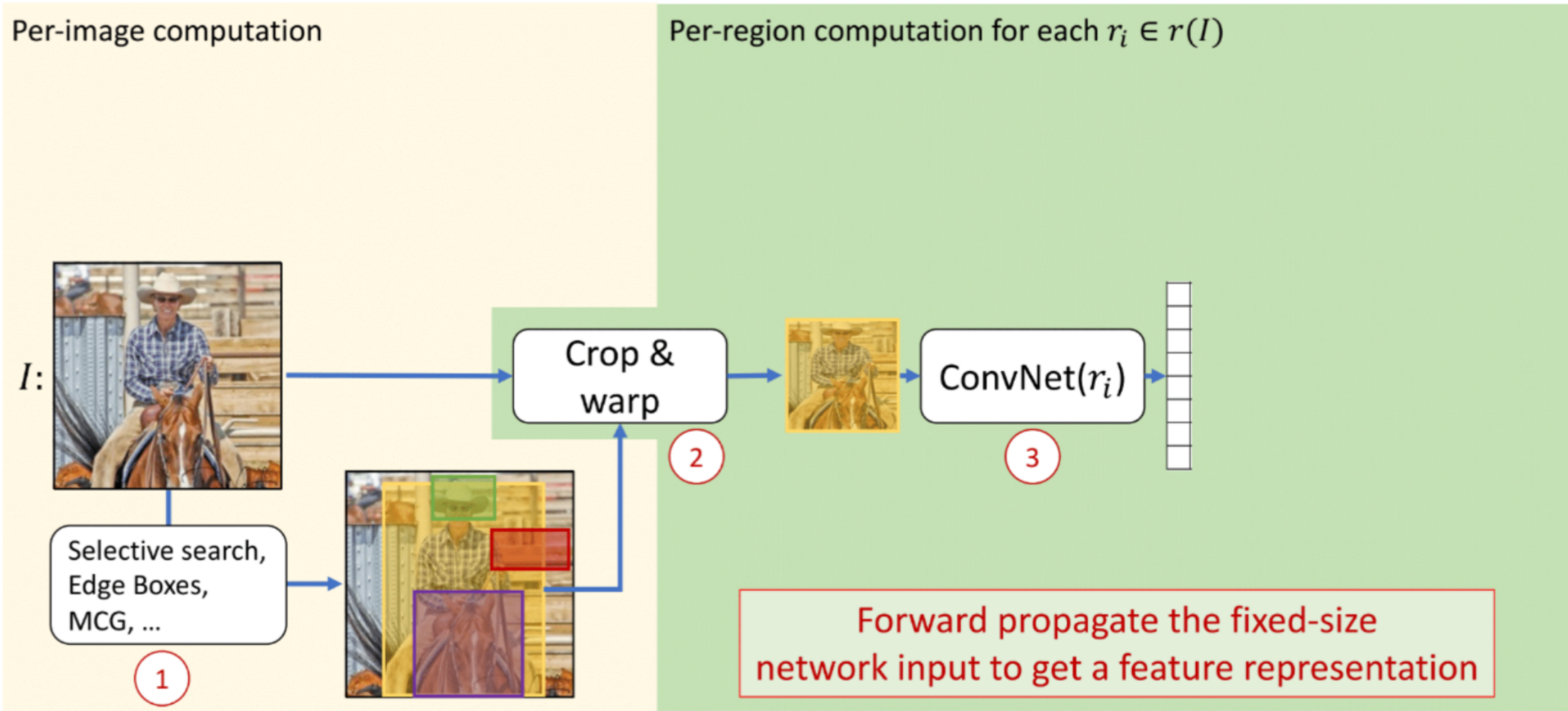
Object Detection Frameworks: R-CNN



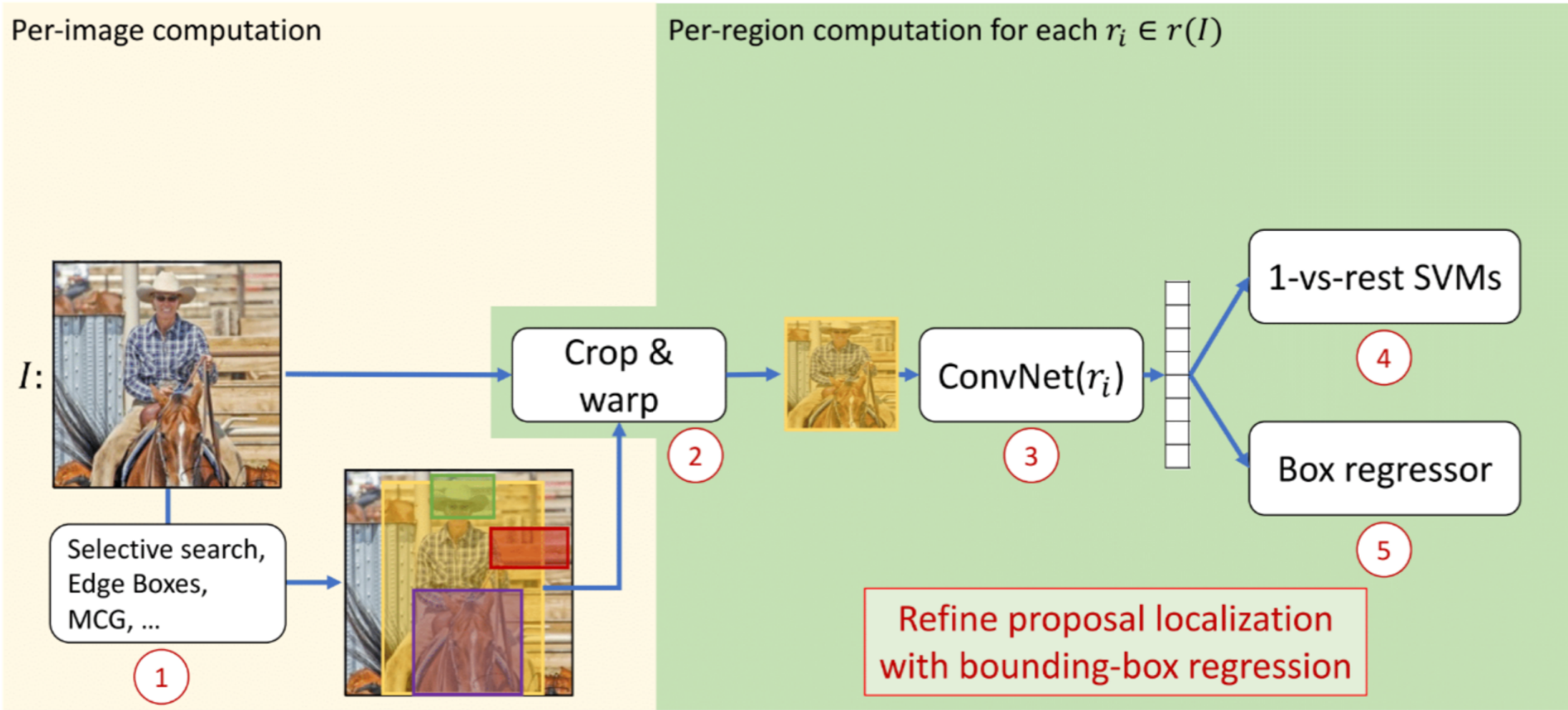
Object Detection Frameworks: R-CNN



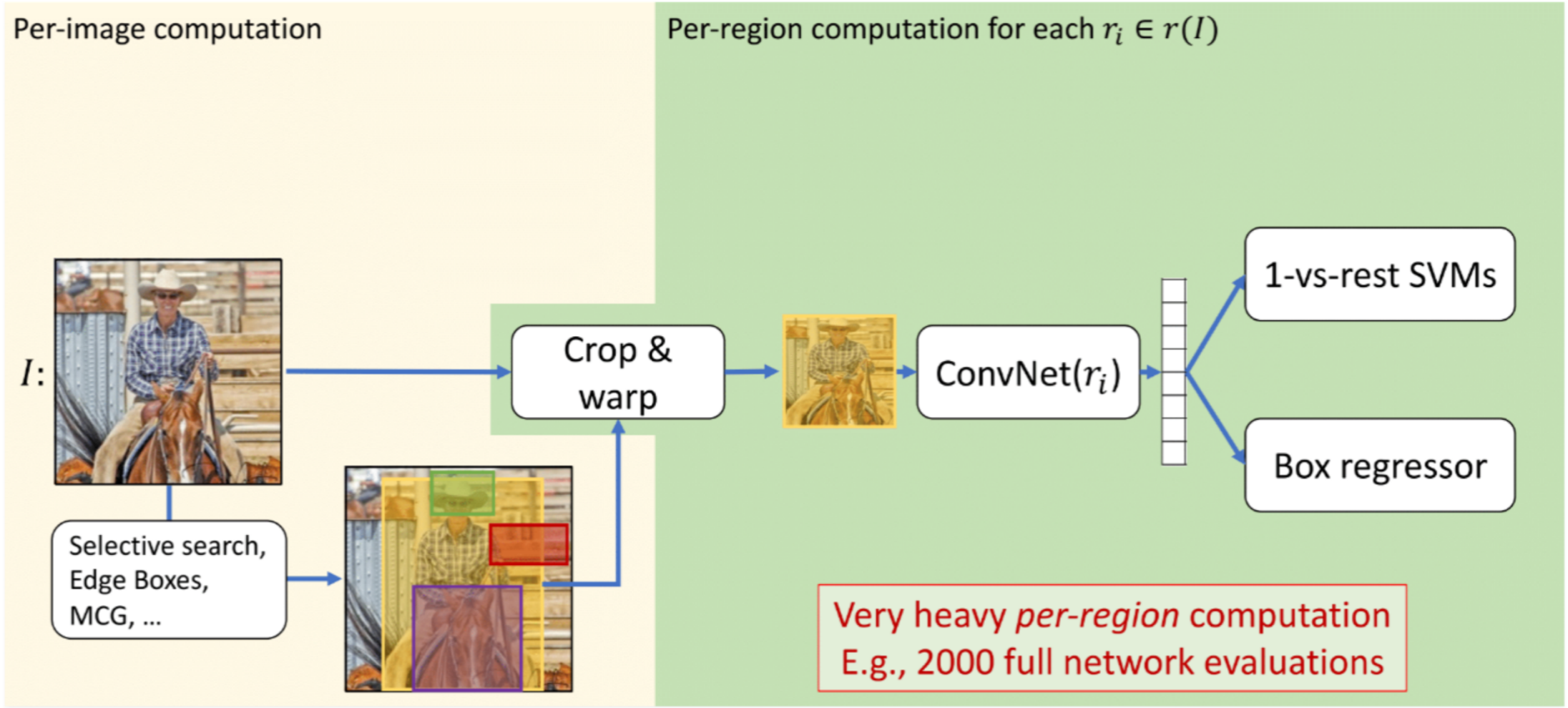
Object Detection Frameworks: R-CNN



Object Detection Frameworks: R-CNN



Object Detection Frameworks: R-CNN



Object Detection Frameworks

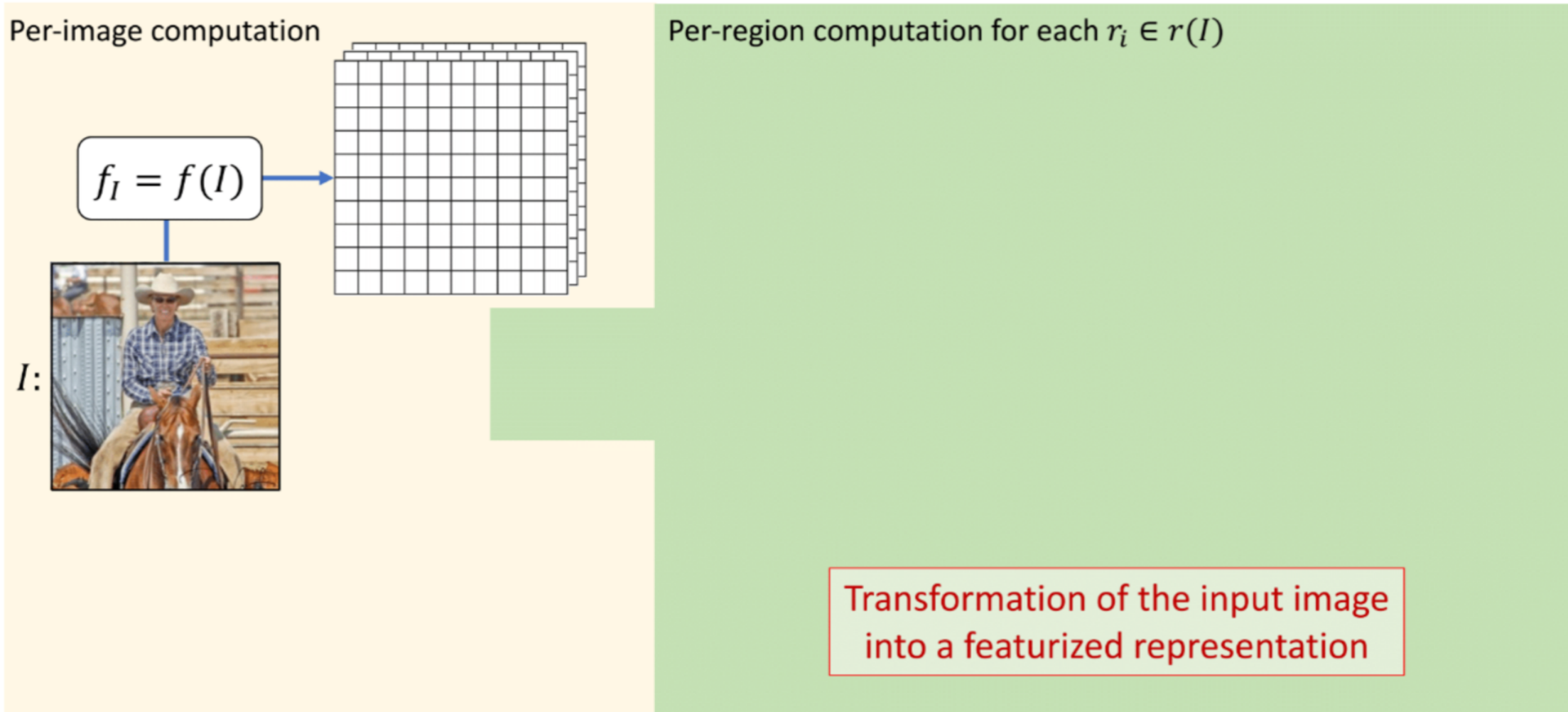
Per-image computation

I:

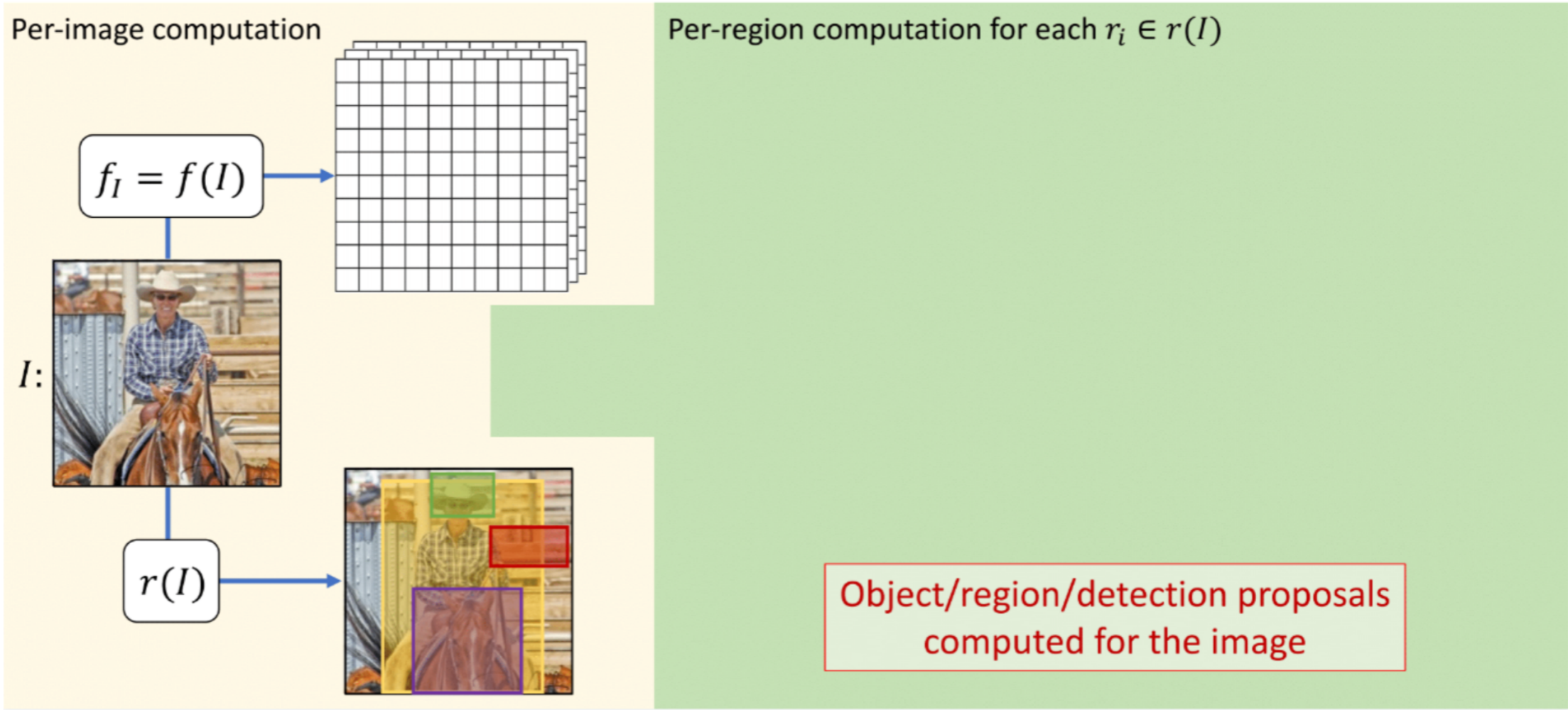
Per-region computation for each $r_i \in r(I)$

Input image
per-image operations | per-region operations

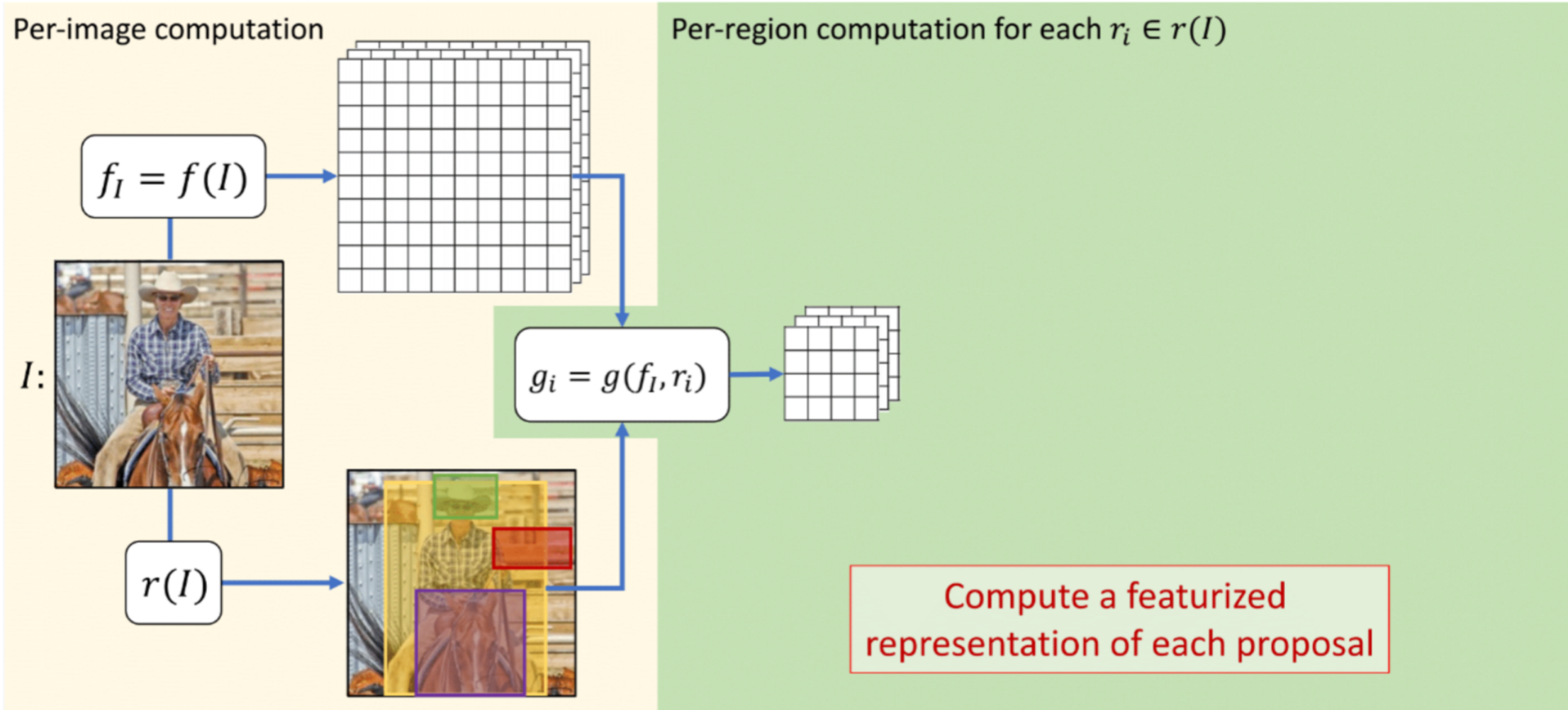
Object Detection Frameworks



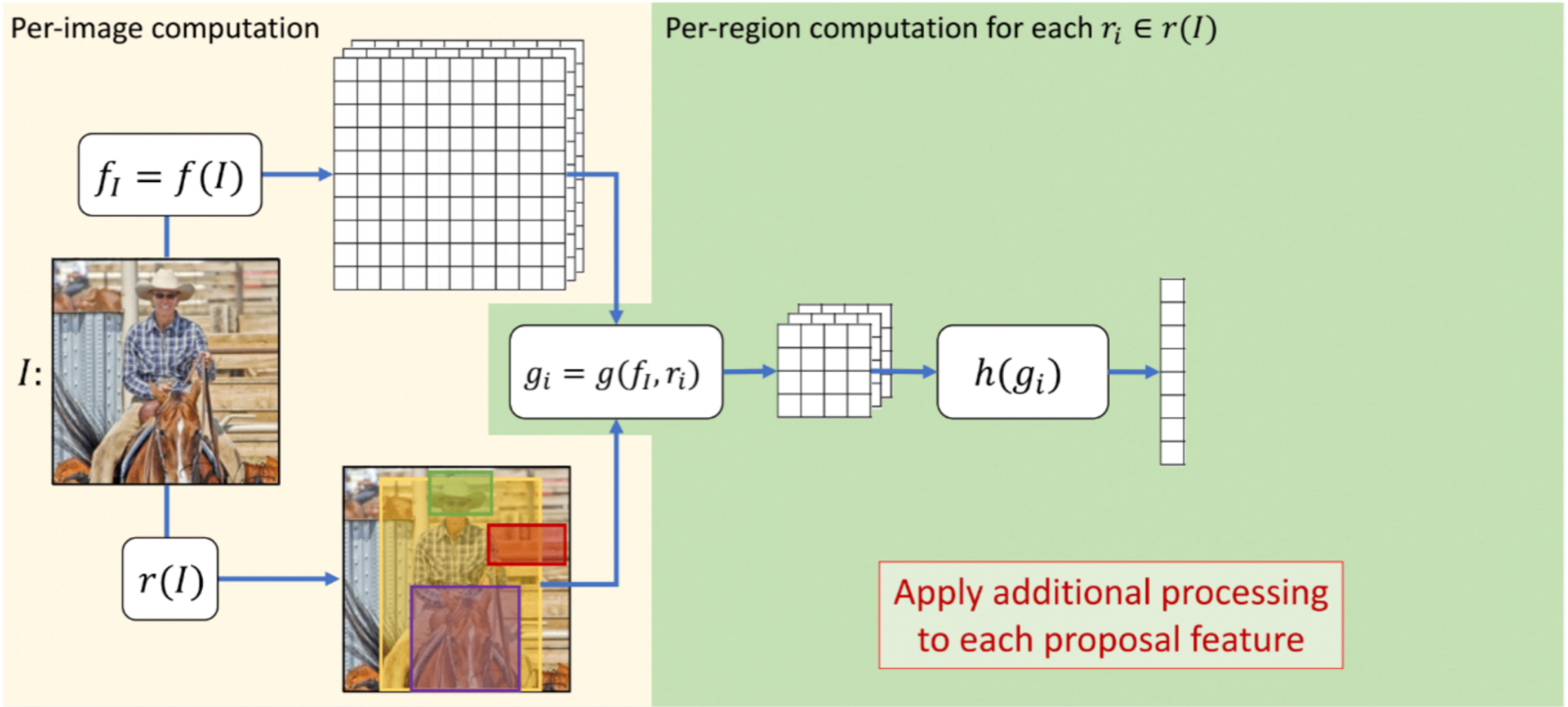
Object Detection Frameworks



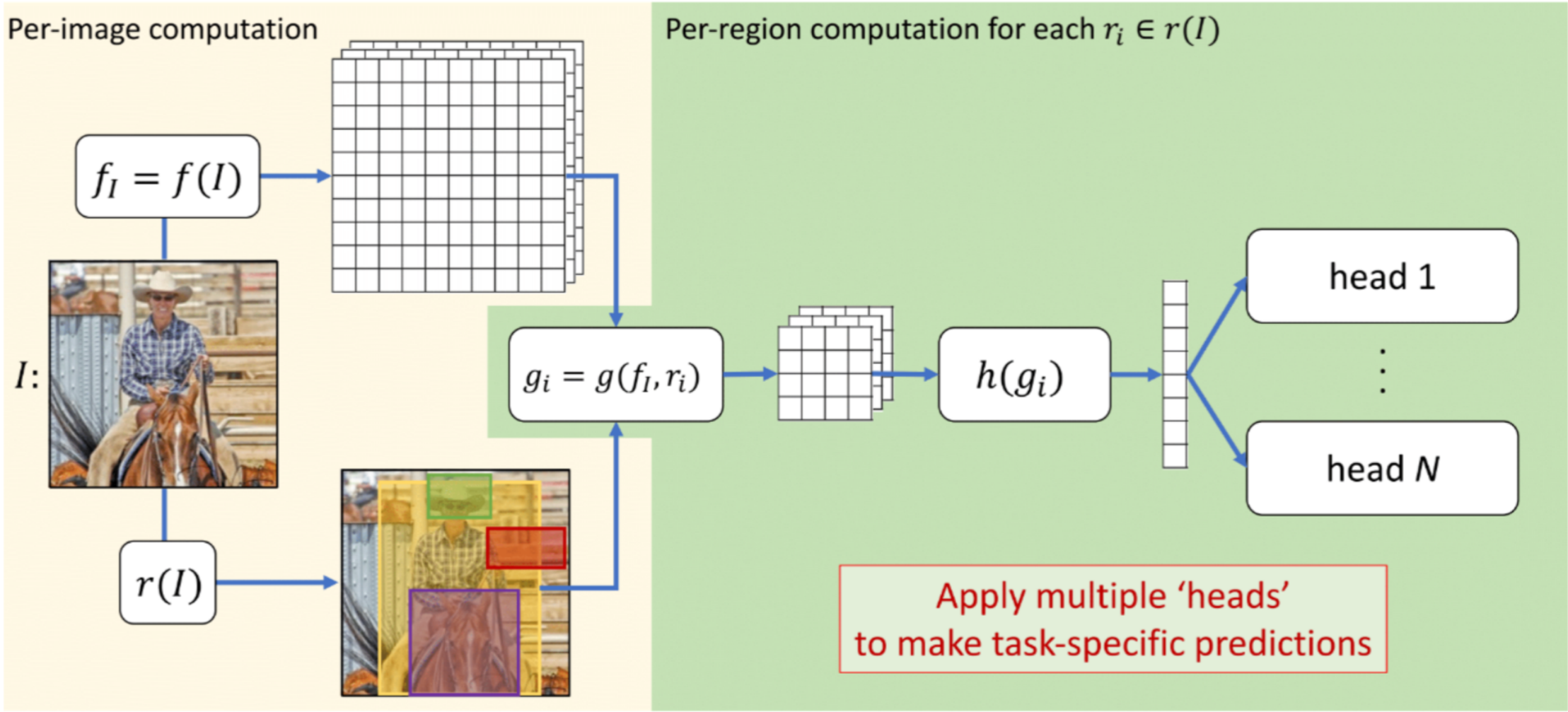
Object Detection Frameworks



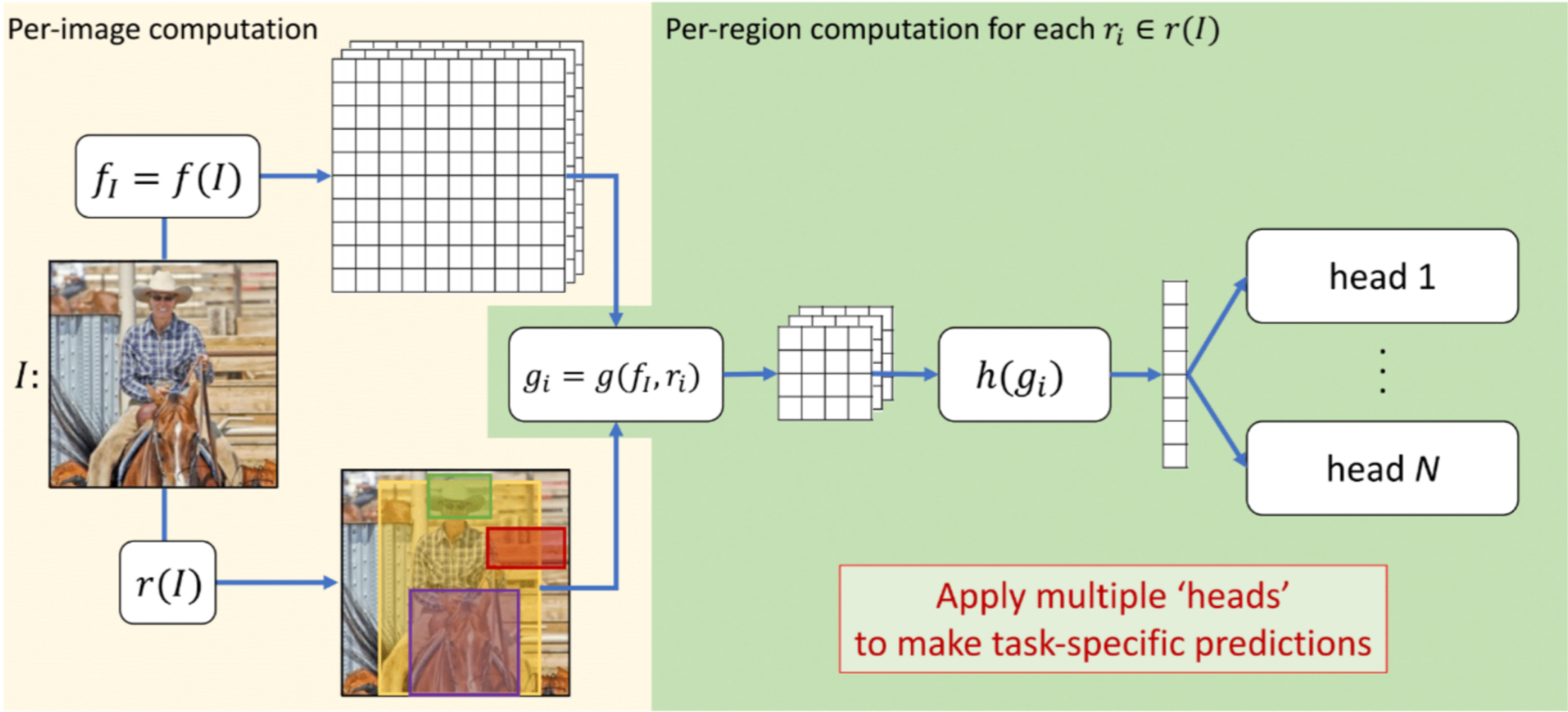
Object Detection Frameworks



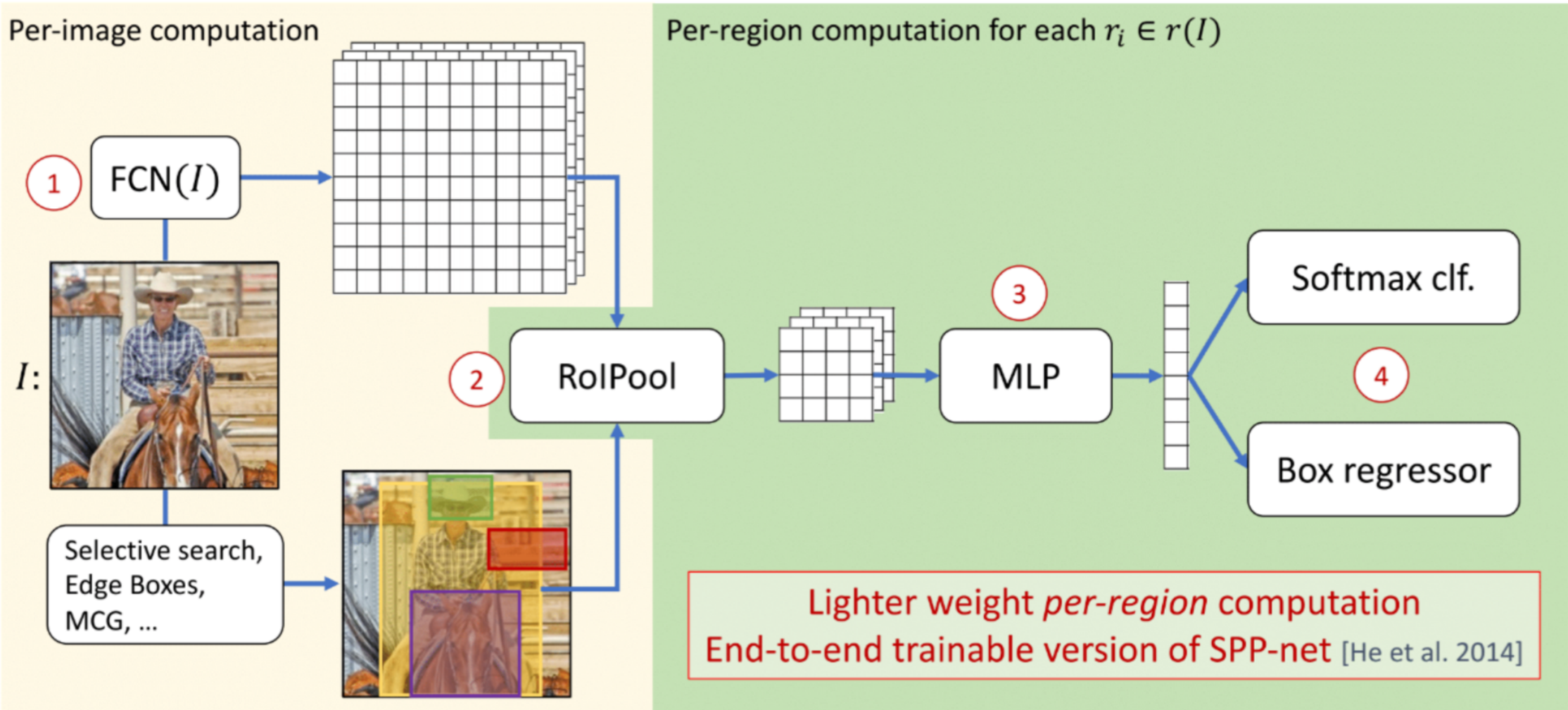
Object Detection Frameworks



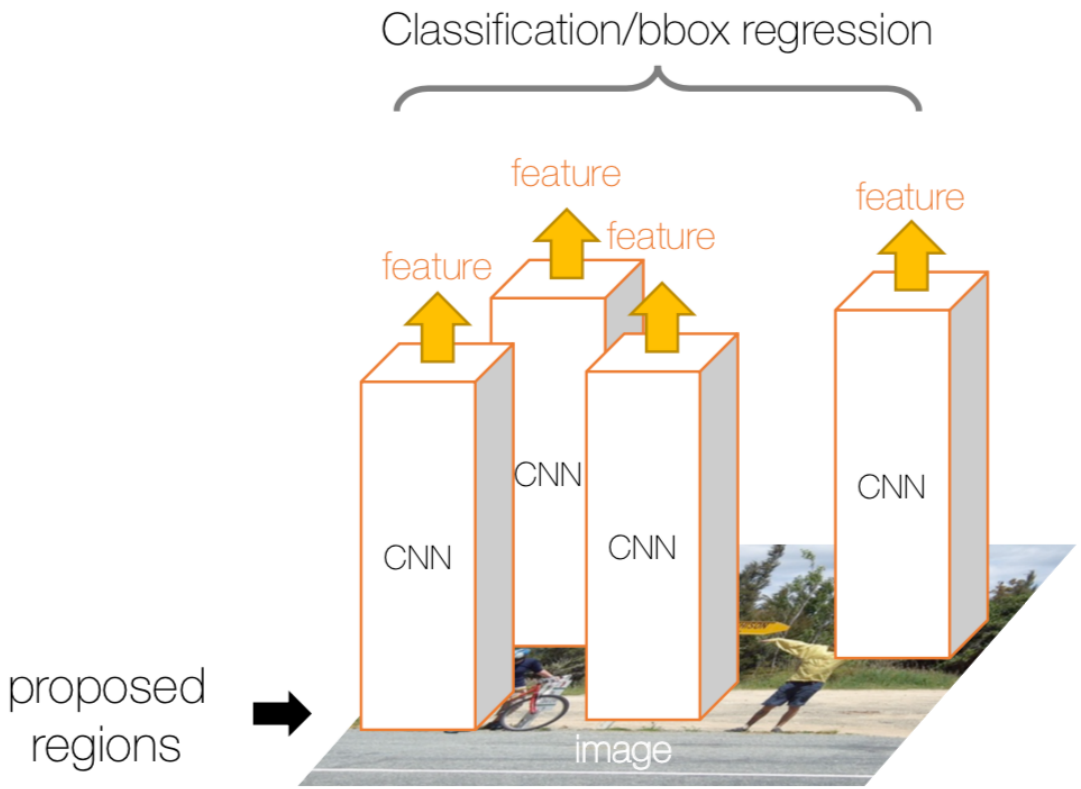
Object Detection Frameworks



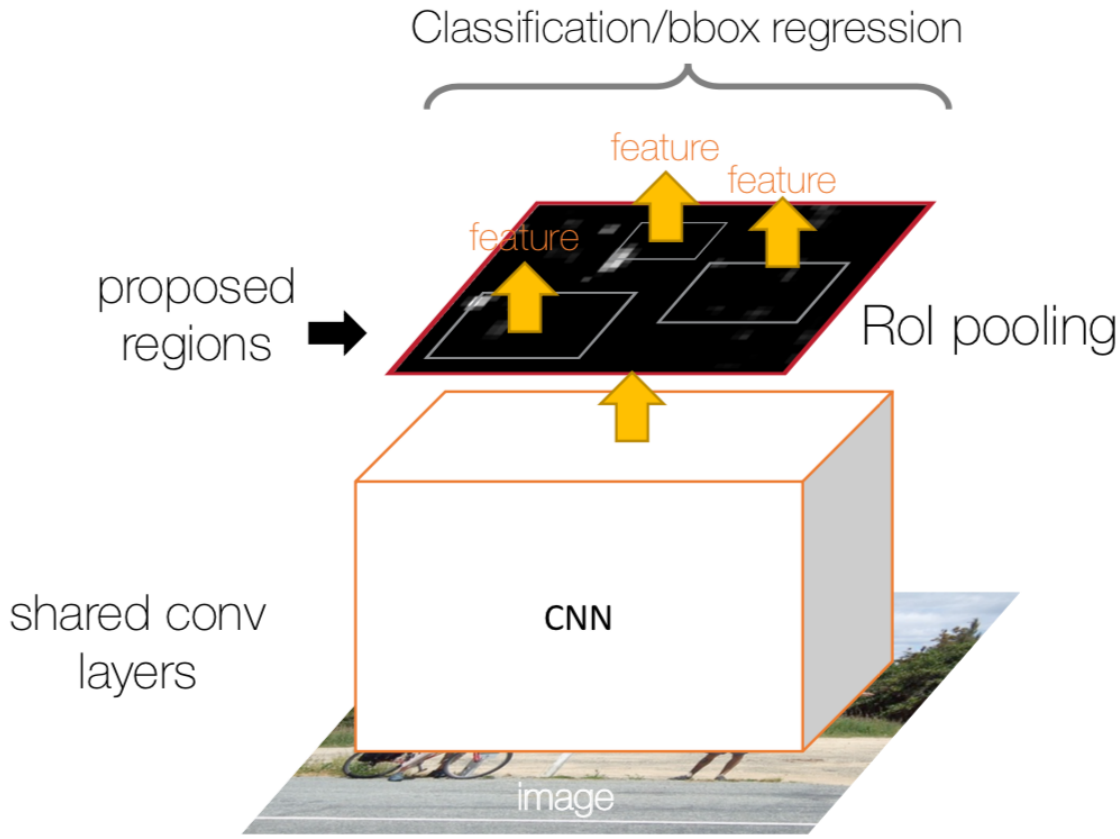
Object Detection Frameworks: Fast R-CNN



Object Detection Frameworks



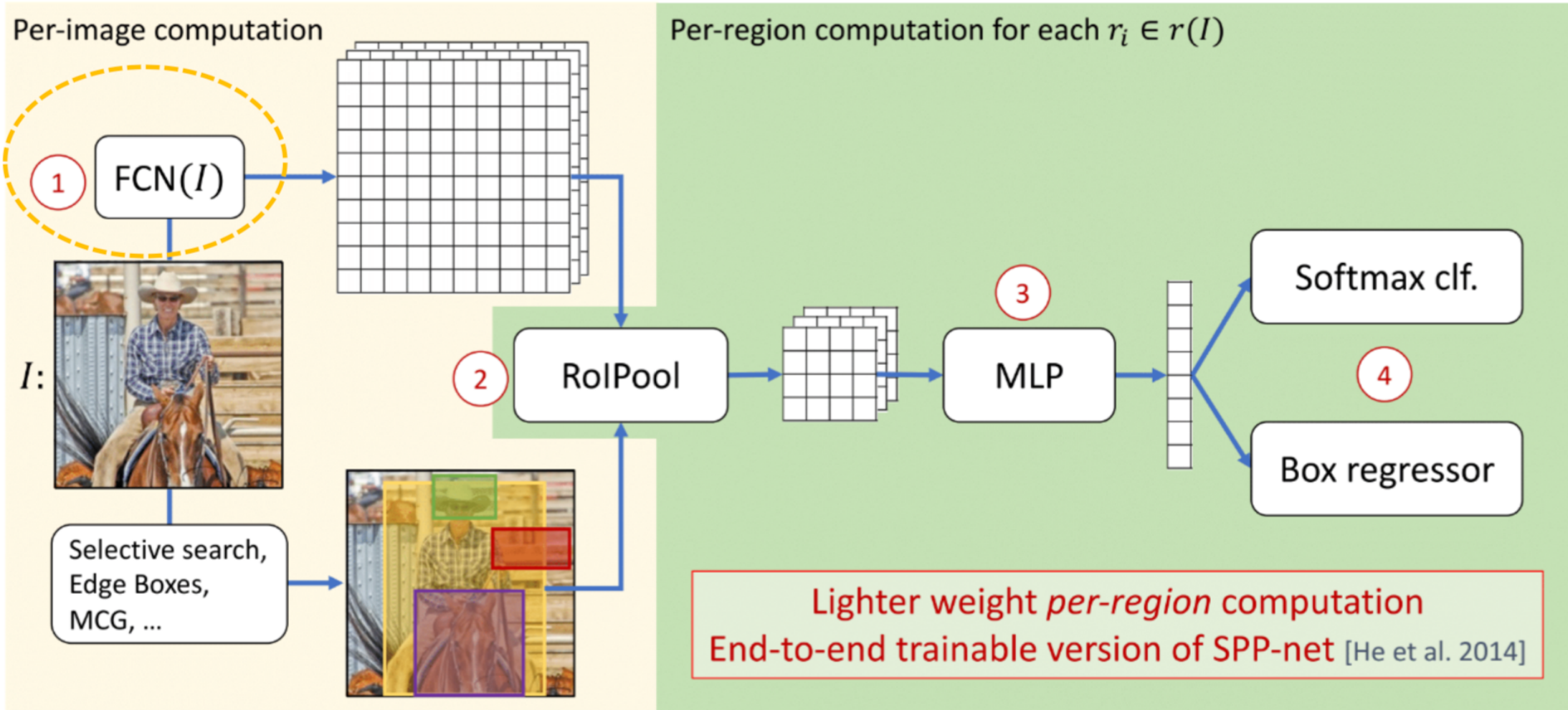
R-CNN



Fast R-CNN

Girshick et al, CVPR 2014 & ICCV 2015

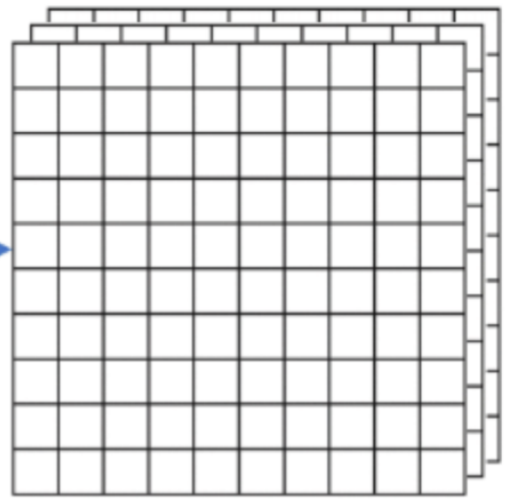
Backbone Networks



Backbone Networks

I:

FCN(I)



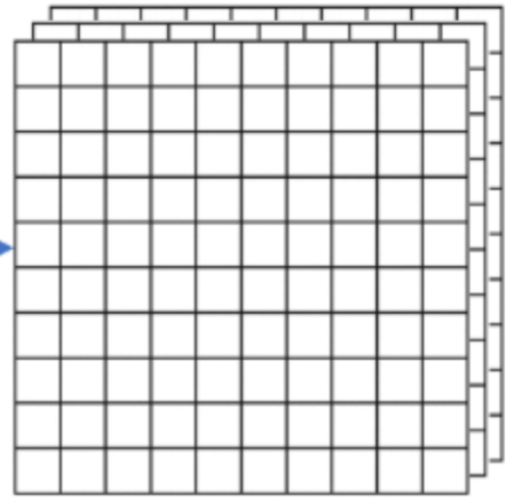
Example feature map dimensions:
(512, H/16, W/16)

Backbone Networks

I:

FCN(I)

AlexNet
VGG16
ResNet
Inception
ResNeXt
DenseNet

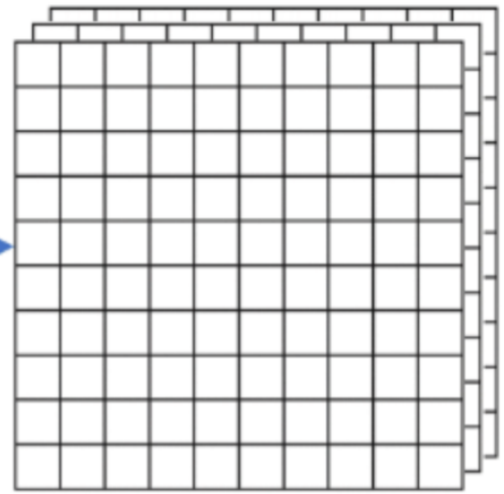


Example feature map dimensions:
(512, H/16, W/16)

Backbone Networks

I: H W

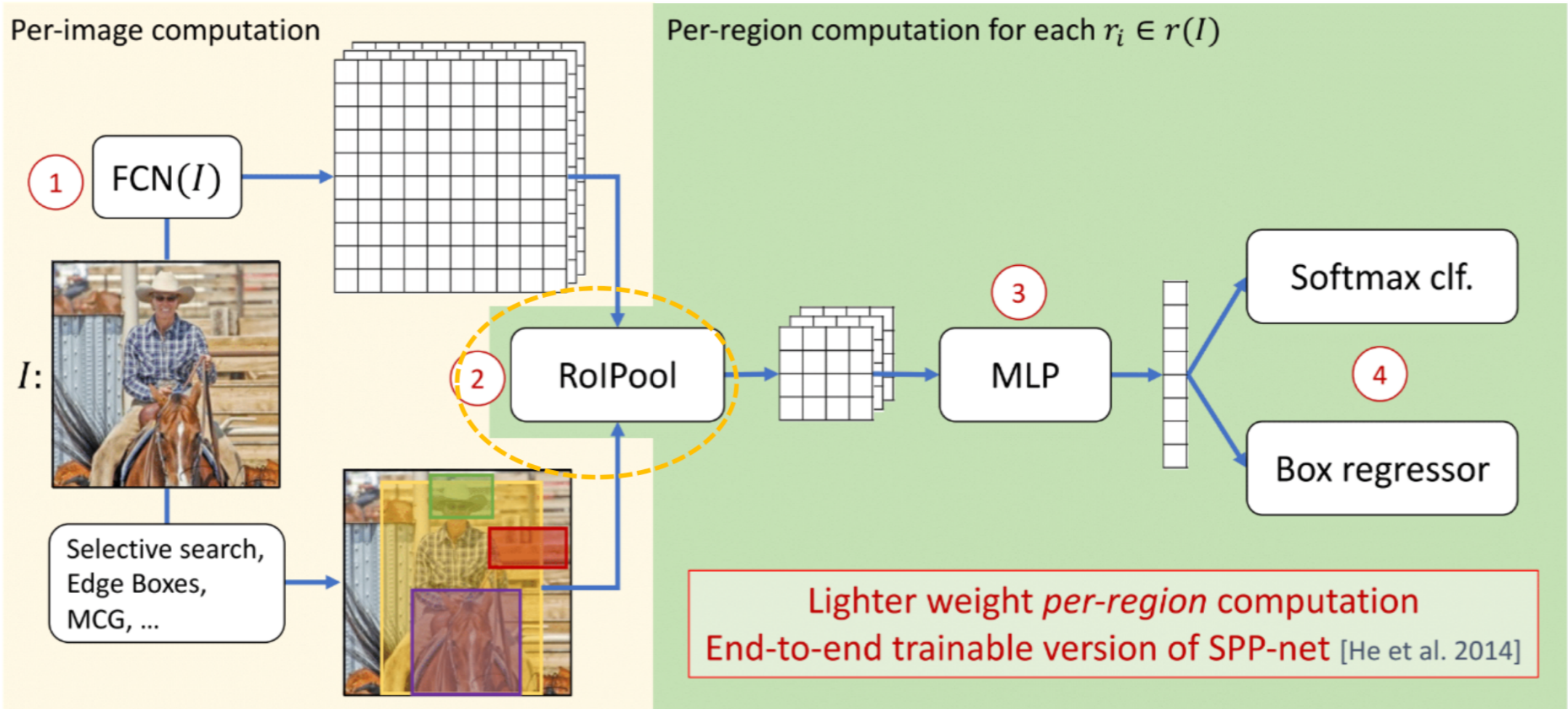
FCN(I)



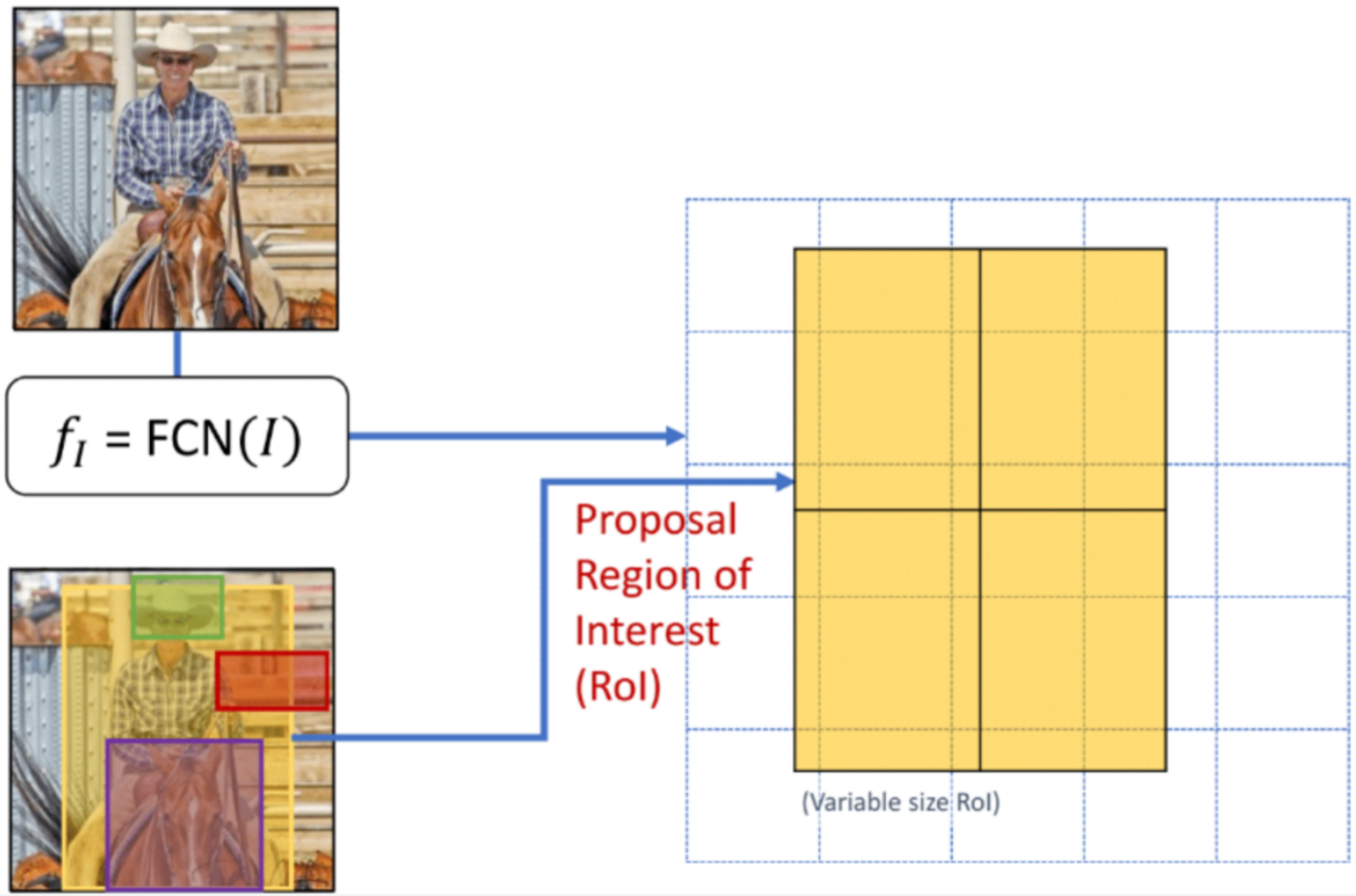
Example feature map dimensions:
(512, H/16, W/16)

Stride of the Backbone: Amount of strided pooling or convolutions

RoIPool

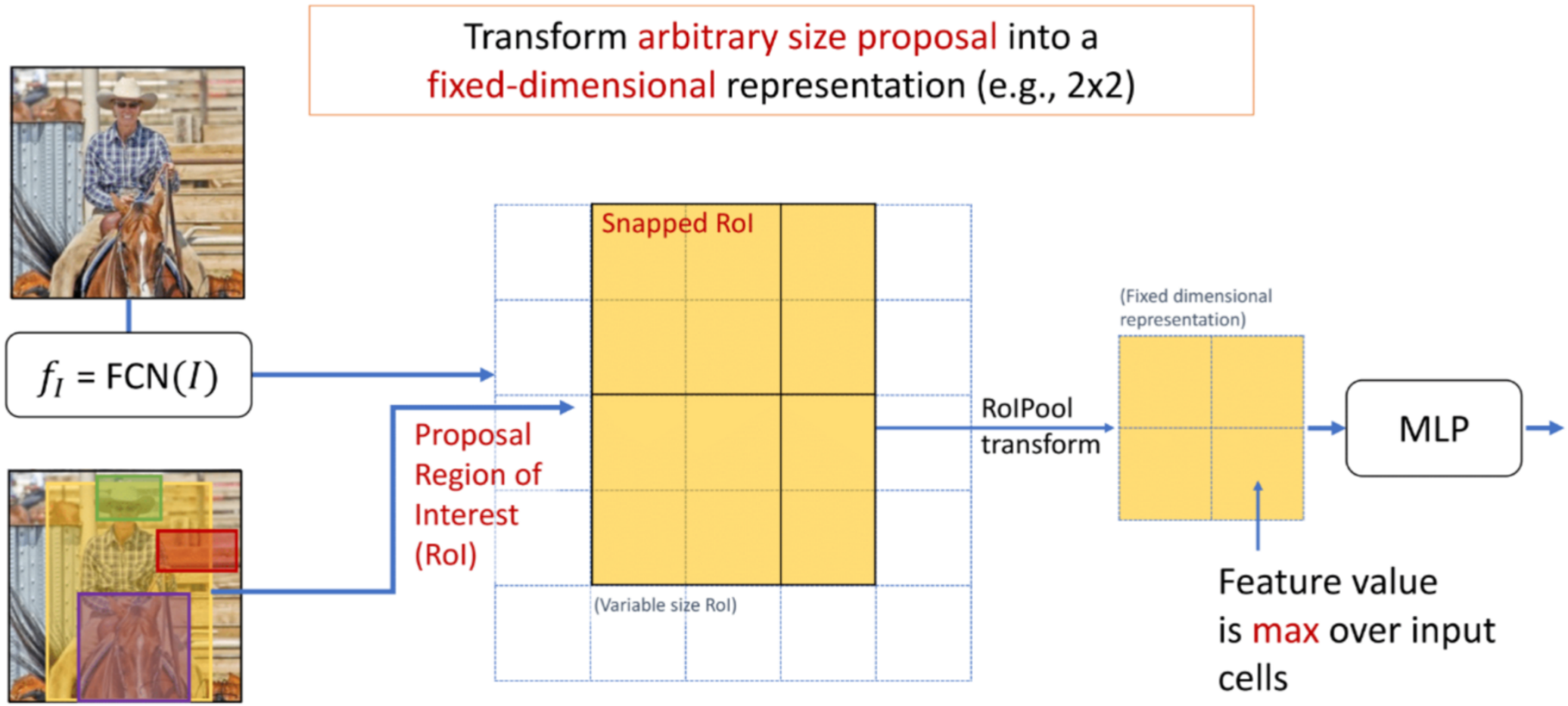

Per-image computation
Per-region computation for each $r_i \in r(I)$
1
FCN(I)
I:
2
RoIPool
3
MLP
4
Softmax clf.
Box regressor
Selective search,
Edge Boxes,
MCG, …
Lighter weight *per-region* computation
End-to-end trainable version of SPP-net [He et al. 2014]

RoIPool on Each Proposal



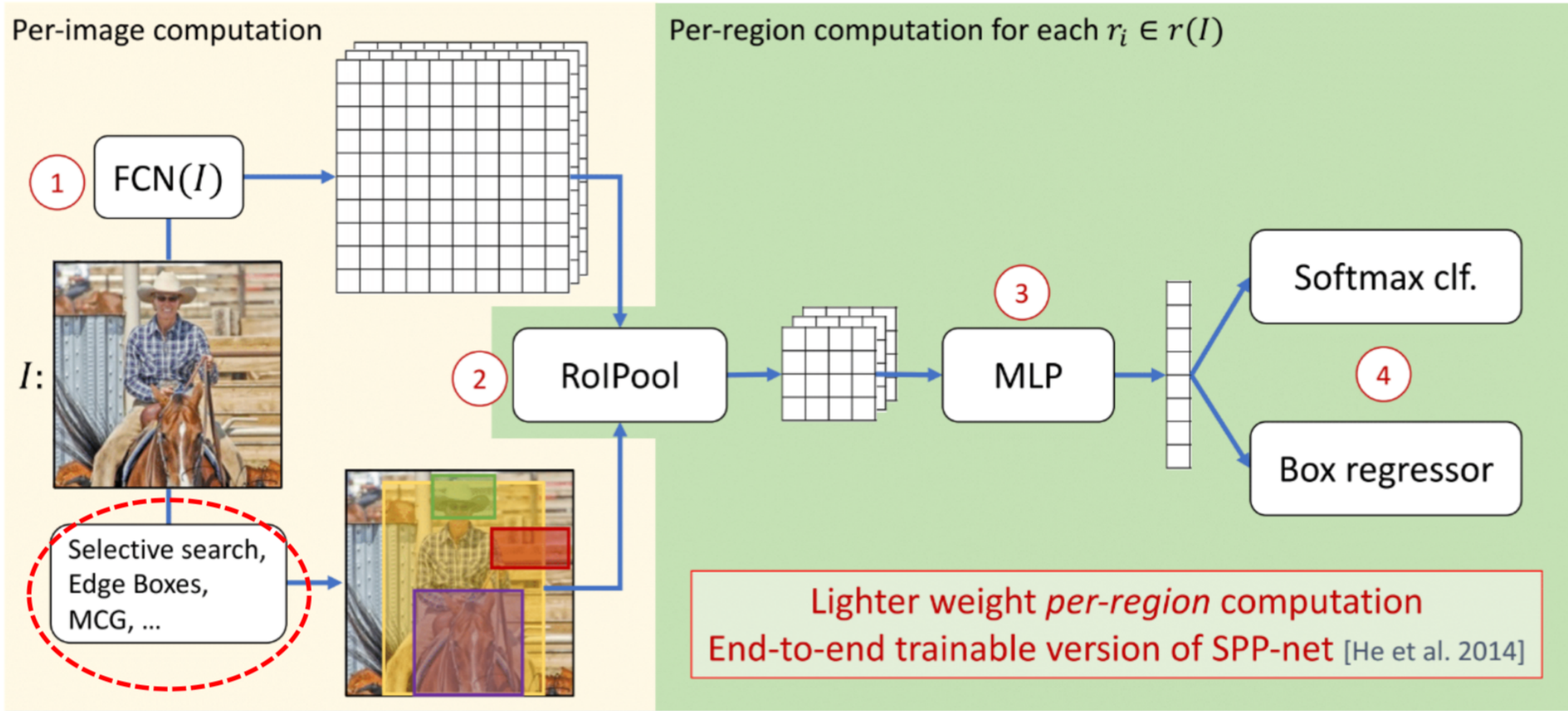
Innovation in SPP-Net [He et al. 2014]

RoIPool on Each Proposal

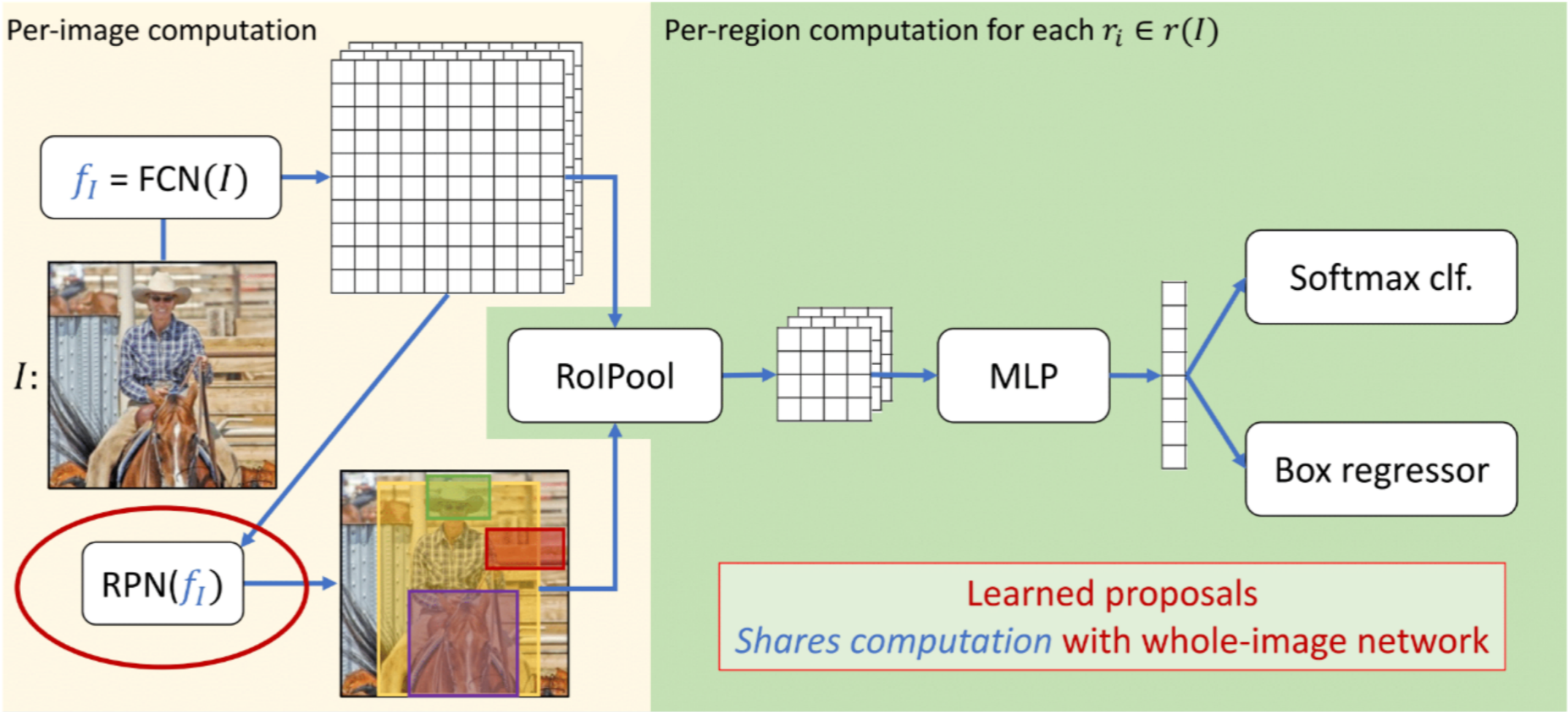


Innovation in SPP-Net [He et al. 2014]

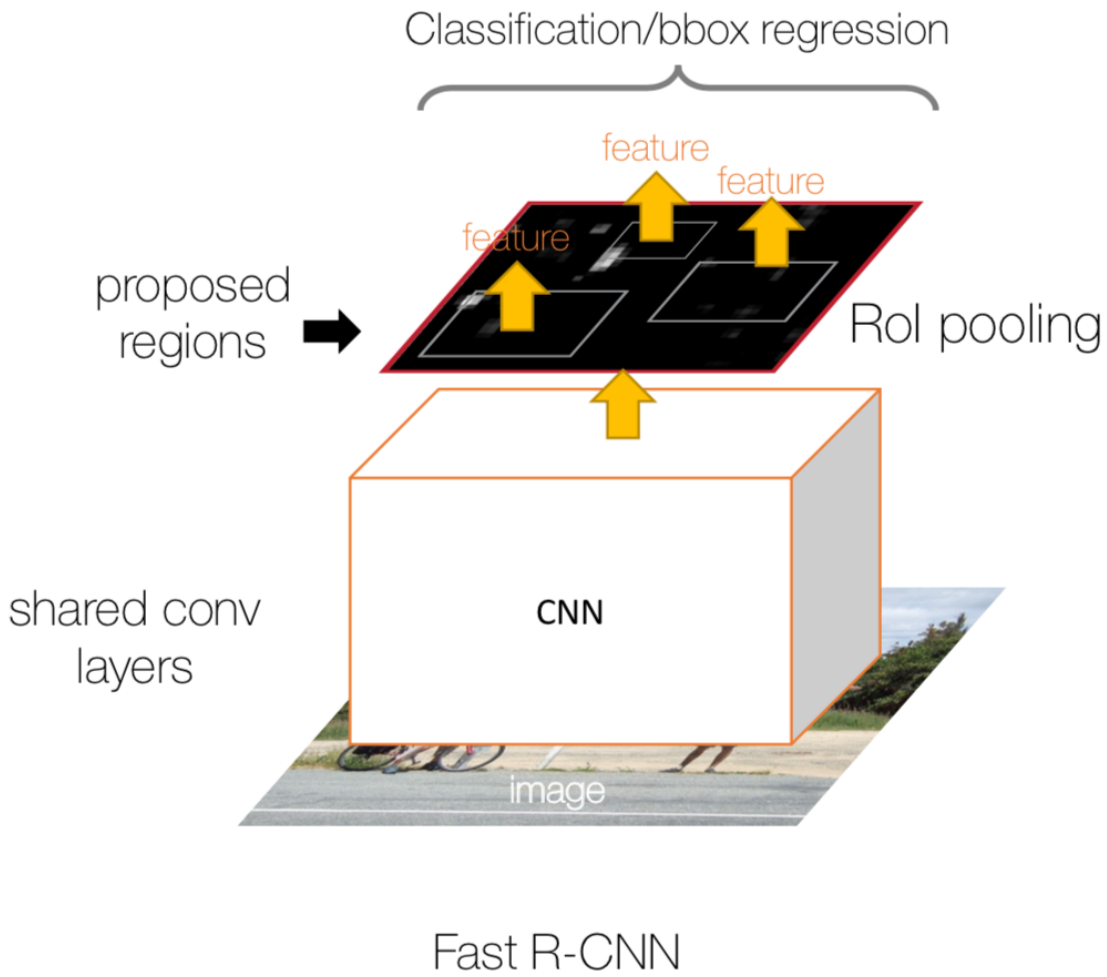
The Problem with Fast R-CNN



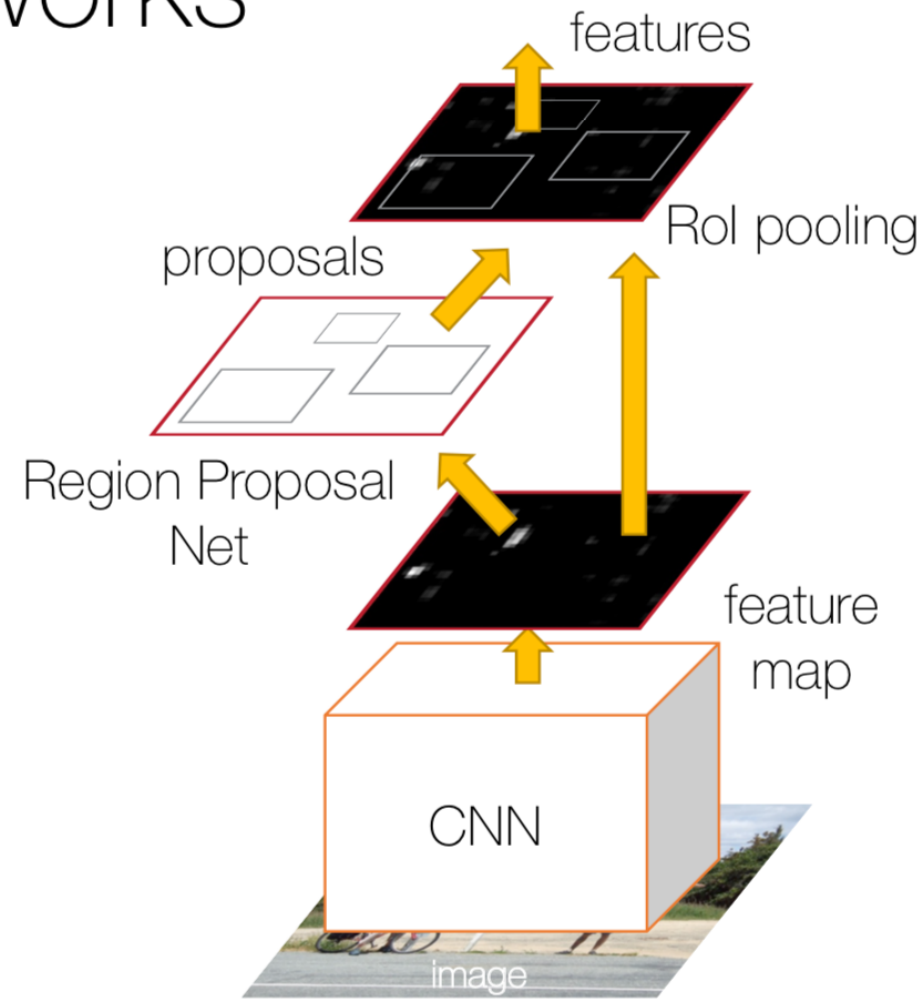
Object Detection Frameworks: Faster R-CNN



Object Detection Frameworks



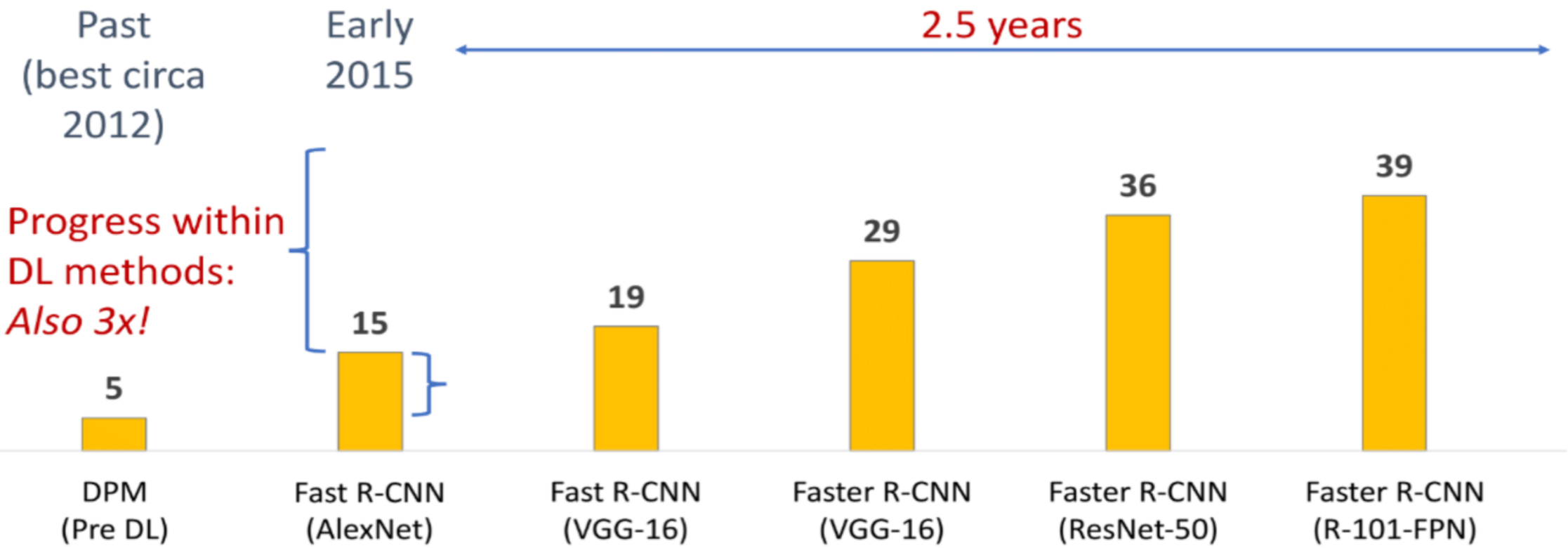
Fast R-CNN



Faster R-CNN

Girshick et al., CVPR 2014 & ICCV 2015, Ren et al., NIPS 2015

Performance through Better Detector Design and Better Backbones



Instance Segmentation

Task

- Assume C object classes
- Segment all object instances in an image

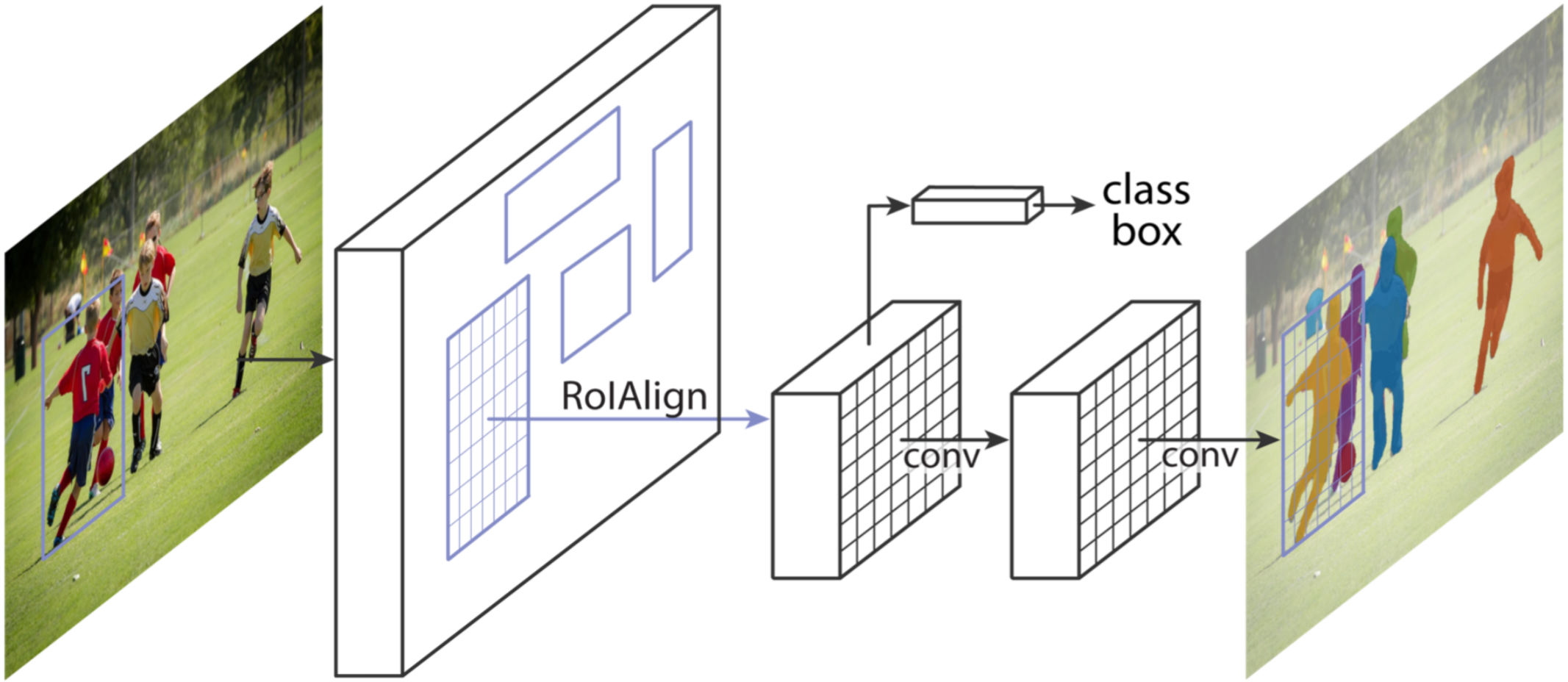
True Positives

- A detection m is a true positive if maskov(m, gt) > 0.5 and class(m) = class(gt)

False Positives

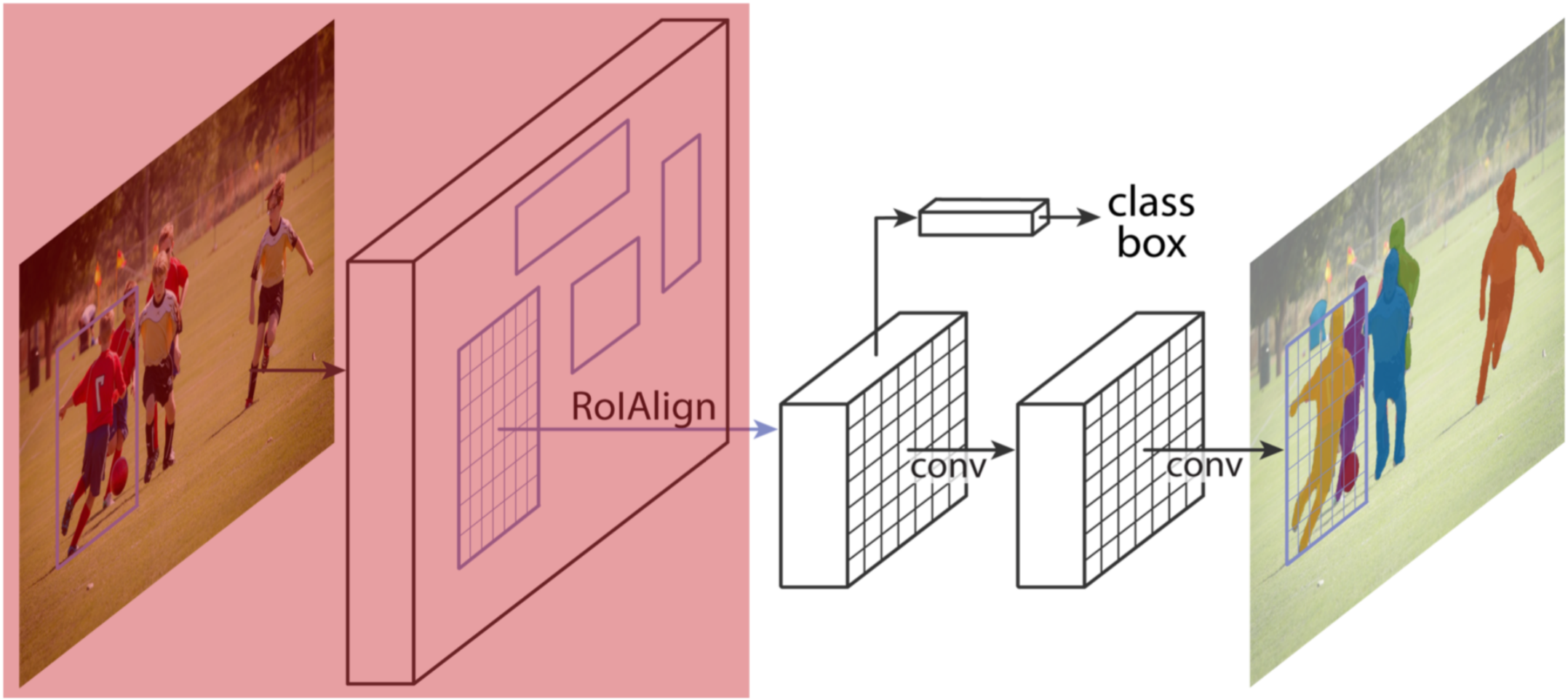
- Mislocalized and misclassified detections
- Duplicate detections are penalized

Mask R-CNN



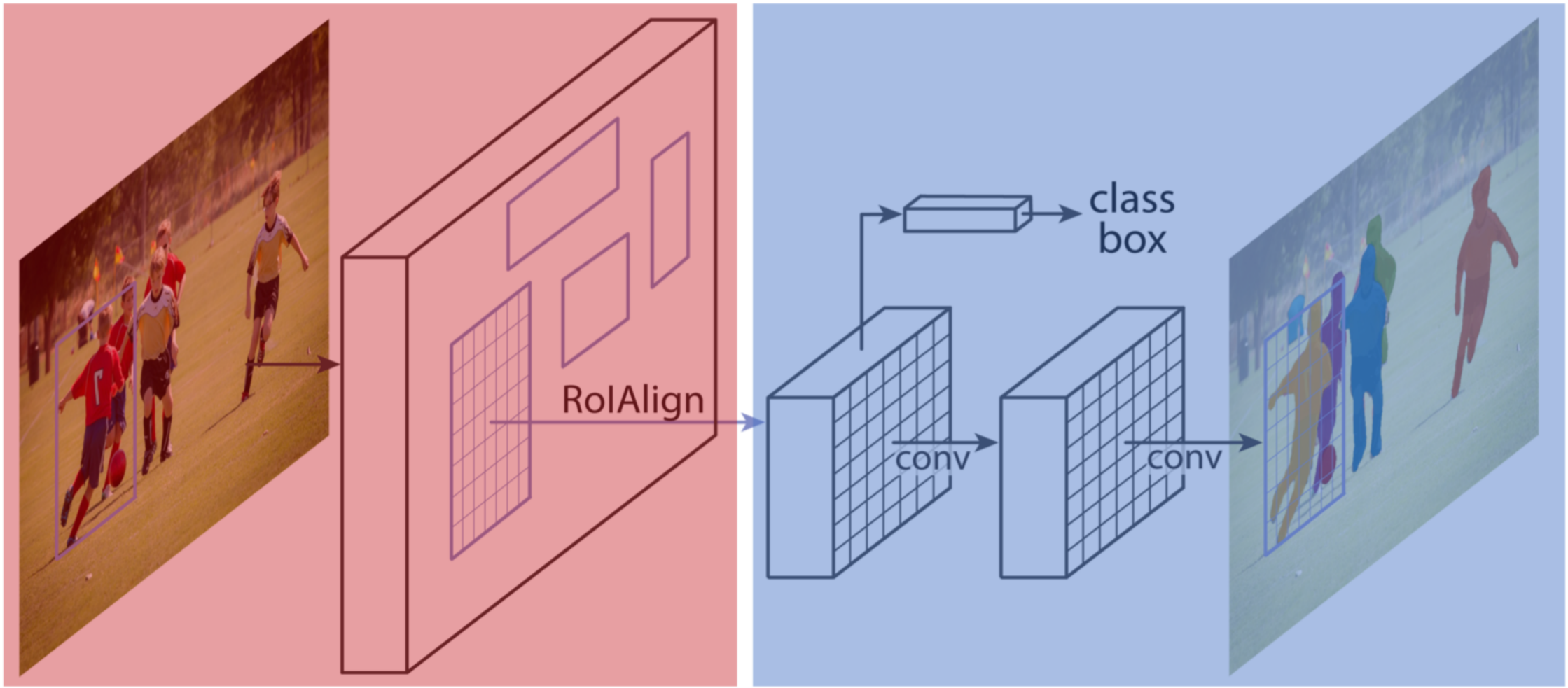
He, Gkioxari, Dollar, Girshick, ICCV 2017

Mask R-CNN



CNN applied to
entire image

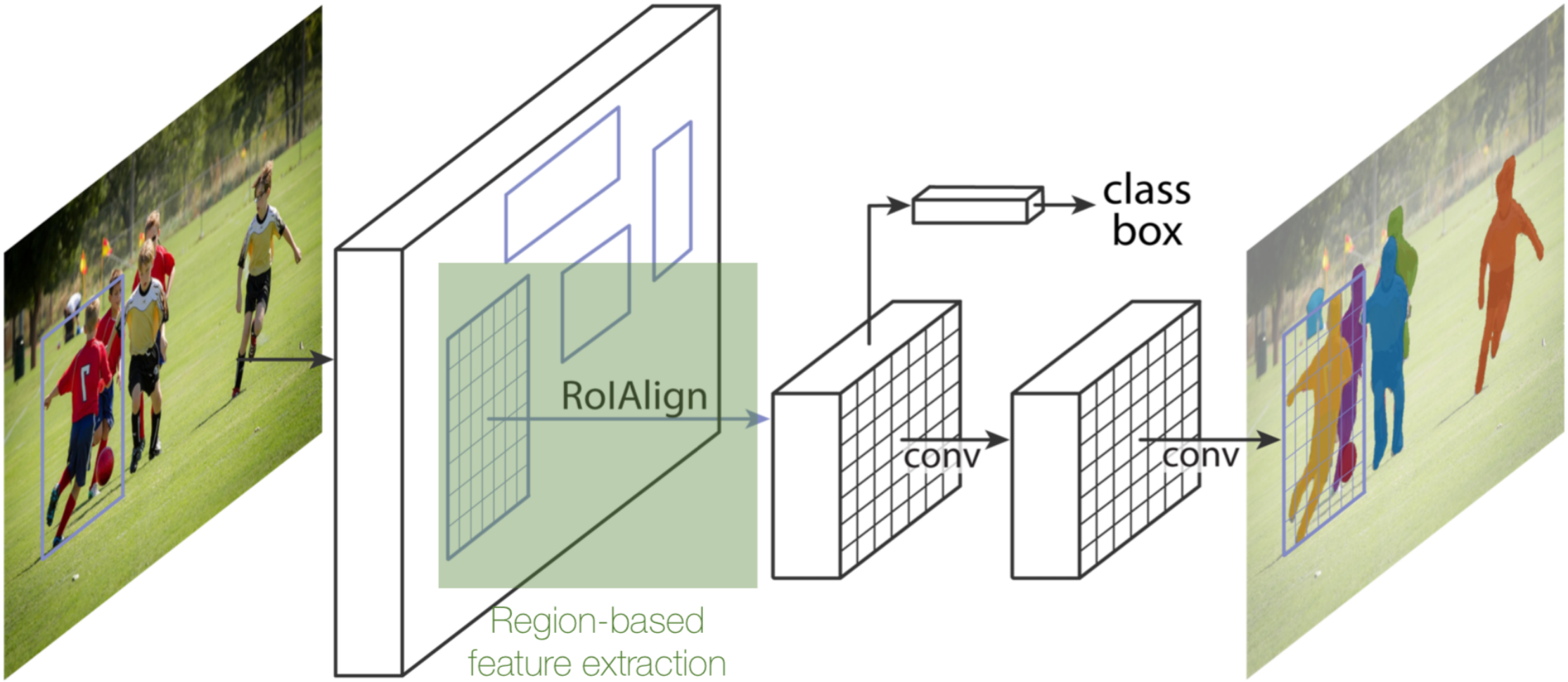
Mask R-CNN



CNN applied to entire image

Region-wise segmentation subnetwork

RoIAlign


class
box
RoIAlign
conv
conv
Region-based
feature extraction

Mask R-CNN for Pose Prediction

Task

Given an image:

- Detect all human instances in the image
- Localize their landmarks, e.g. Right Shoulder, Nose etc.

Mask R-CNN for Pose Prediction

Loss

- Assume a set of K landmarks, e.g. Nose, Right Shoulder etc.
- For a detected instance R
  - corresponding logits J predicted by the Pose Head of Mask R-CNN of shape H x W x K
  - Ground truth locations $J_{gt} = \{(x_k, y_k)$, for k = 0, …, K-1$\}$

Mask R-CNN for Pose Prediction

Loss

- Assume a set of K landmarks, e.g. Nose, Right Shoulder etc.
- For a detected instance R
  - corresponding logits J predicted by the Pose Head of Mask R-CNN of shape H x W x K
  - Ground truth locations $J_{gt} = \{(x_k, y_k)$, for k = 0, …, K-1$\}$

References

- Georgia Gkioxari, Object Recognition slides AMMI 2019.
- Laurens van der Maaten and Anton Bakhtin, introduction to convnets AMMI 2019.