Object Recognition and Segmentation

From R-CNN to Mask R-CNN

Daniel Eftekhari

daniel.eftekhari.ai@gmail.com

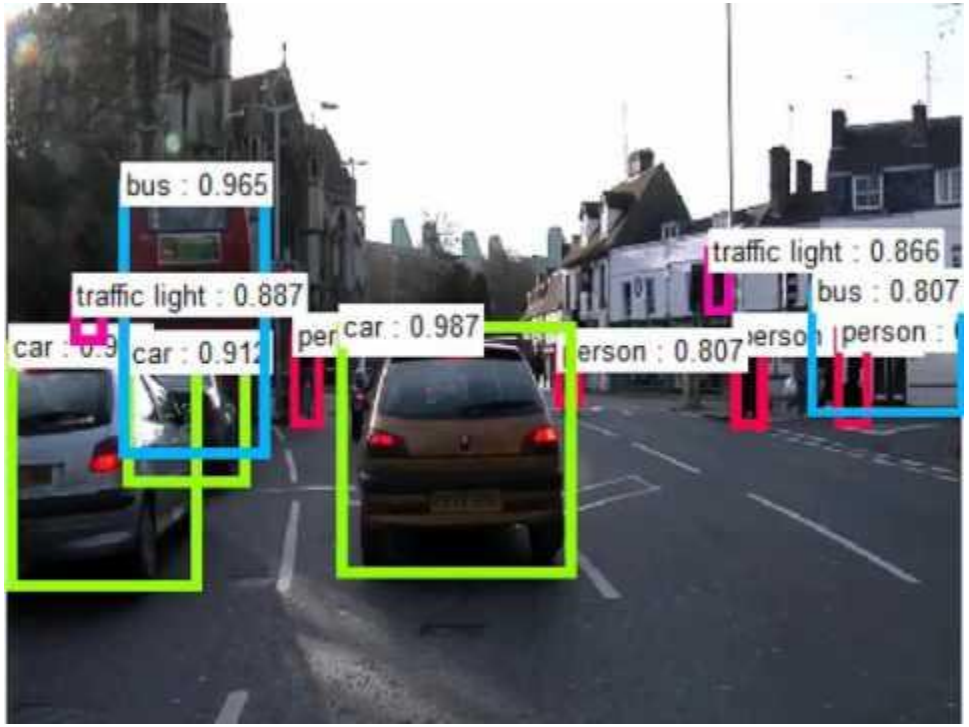
Motivation

- Convolutional neural networks (CNNs)
  - Reliable tools for image classification

- But…

To what extent do [Krizhevsky et. al's results] generalize to object detection?

- Girshick et al.


bus : 0.965
traffic light : 0.887
car : 0.912
car : 0.987
traffic light : 0.866
bus : 0.807
person :

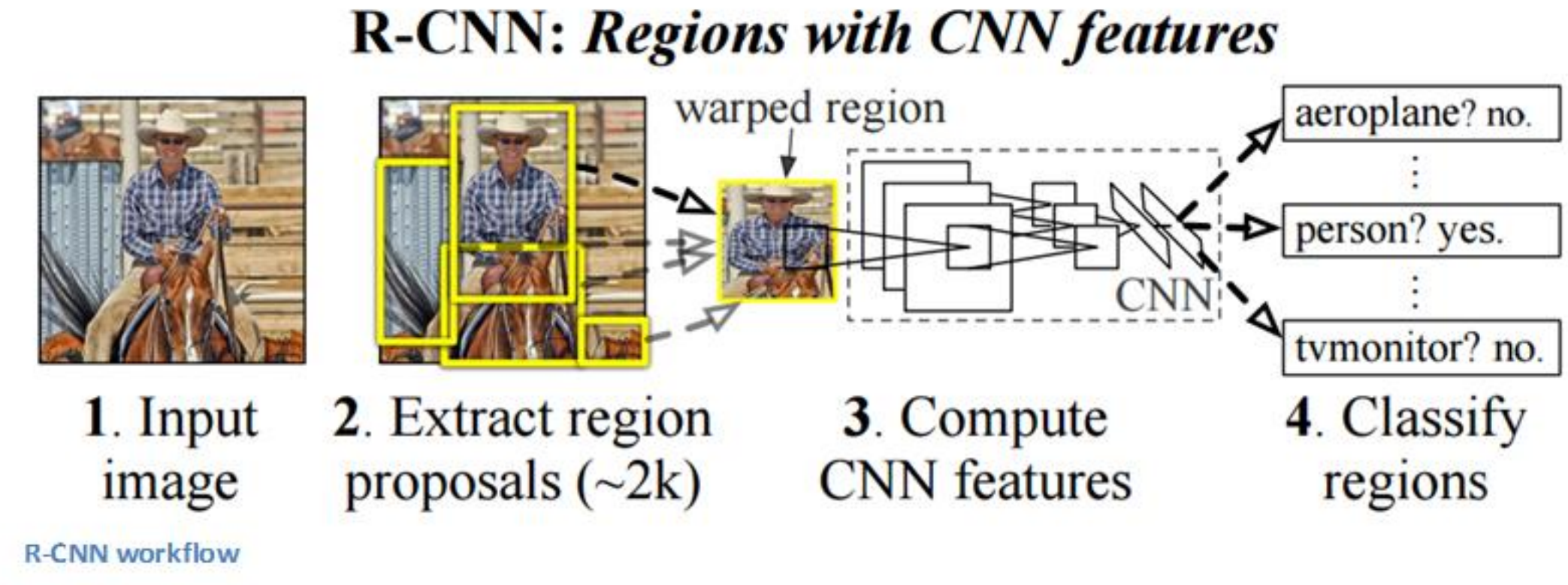
Regional CNN (R-CNN) [1]

- Goal: Generate bounding boxes for objects and produce labels

- Three tasks
  - Region proposal
    - Selective search [2] & Affine image warping
  - Classification
    - CNN + SVM
  - Tightening bounding boxes
    - Bounding-box regression

[1] Girshick et al. (2014), https://arxiv.org/abs/1311.2524
[2] Uijlings et al. (2012), https://link.springer.com/article/10.1007/s11263-013-0620-5

R-CNN overview



Girshick et al. (2014), https://arxiv.org/abs/1311.2524

Selective search & Affine warping

- Initial segmentation using a graph-based segmentation method [1]

- Regions are proposed in a bottom-up fashion using similarity in
  - Color (histogram intersection after normalizing color histograms)
  - Texture (histogram intersection of Gaussian derivatives at 8 orientations)
  - Size (encourages small regions to merge)
  - Fill (fills gaps between regions if they form a natural boundary)

- Proposed regions are transformed using affine warping and fed into a standard CNN, which in turn extracts a feature vector for each region

[1] Felzenszwalb et al. http://cs.brown.edu/~pff/papers/seg-ijcv.pdf

CNN & SVM layers

- Pre-trained (on ImageNet) CNN tuned using warped images
  - Extracts fixed-length feature vector for each region

- Feature vector fed into a class-specific SVM
  - No theoretical justification for this
  - 21-way softmax regression classifier can be used

Bounding box regression

- Transform proposed region P to ground-truth box G
  - $(P_x, P_y, P_w, P_h) \rightarrow (G_x, G_y, G_w, G_h)$
    - coordinates for image center (x, y)
    - width and height (w, h)

- Ridge regression was used to find the mapping

$$\hat{G}_x = P_w d_x(P) + P_x$$
$$\hat{G}_y = P_h d_y(P) + P_y$$
$$\hat{G}_w = P_w \exp(d_w(P))$$
$$\hat{G}_h = P_h \exp(d_h(P)).$$

$$t_x = (G_x - P_x)/P_w$$
$$t_y = (G_y - P_y)/P_h$$
$$t_w = \log(G_w/P_w)$$
$$t_h = \log(G_h/P_h).$$

Evolution of R-CNN

R-CNN → Fast R-CNN → Faster R-CNN

Problems with R-CNN

- Three problems with R-CNN
  - Training is a multi-stage pipeline
    - Gives rise to problems in effectively optimizing the model
  - Training is expensive in space and time
    - For SVM and bounding-box regressor training, features are extracted from each object proposal in each image and written to disk
  - Object detection is slow
    - Forward pass of the CNN required for every single region proposal for every single image

Fast R-CNN [1]

- Training is single-stage, and can update all network layers
  - Jointly train CNN, softmax classifier and bounding box regressor
  - Uses a multi-task loss
- No disk storage is required for feature caching
  - A result of not caching features
- Higher detection quality than R-CNN
  - Eliminated need for SVM
    - Showed softmax classifier can outperform SVM
  - Showed sparse region proposals outperform dense region proposals

[1] Girshick (2015), https://arxiv.org/abs/1504.08083

Fast R-CNN overview

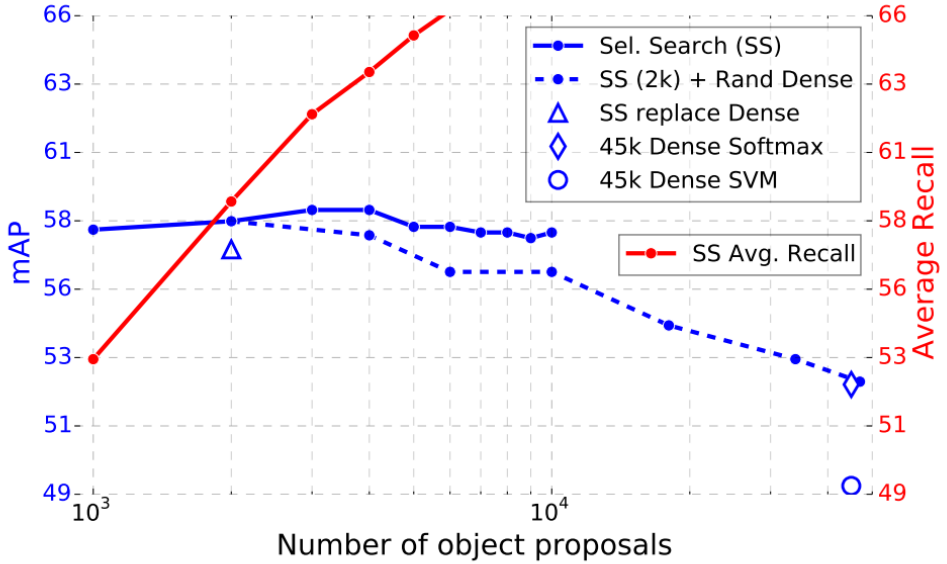
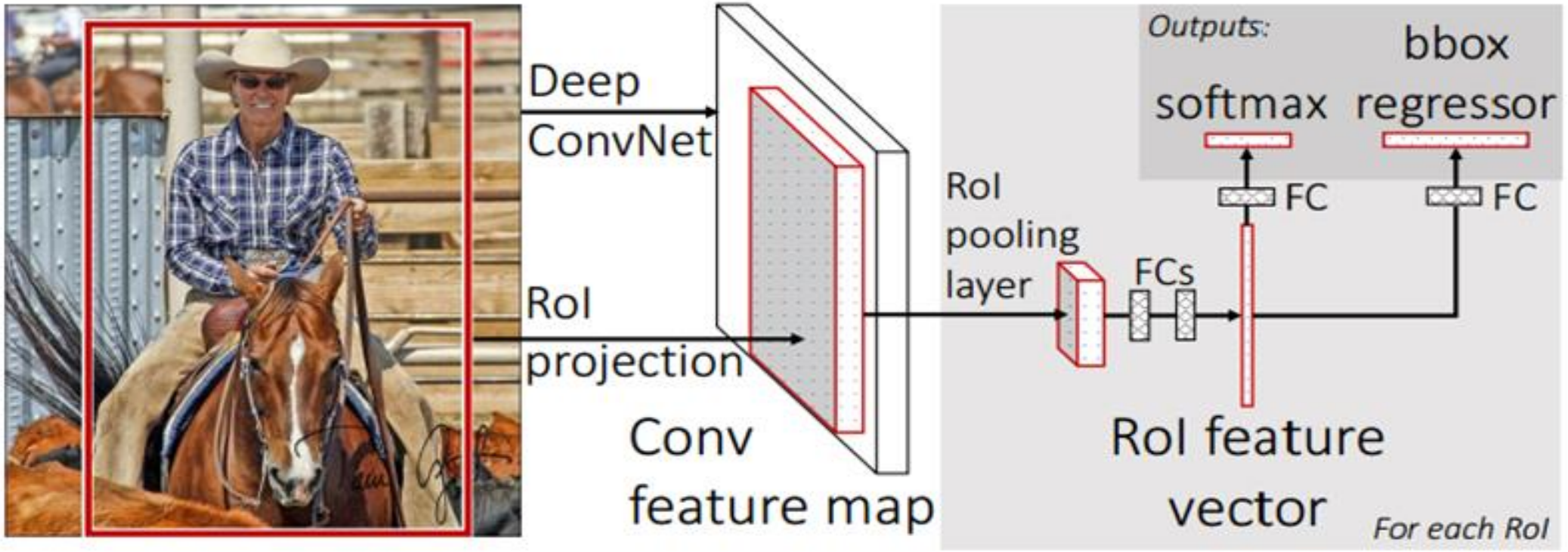


Figure 3. VOC07 test mAP and AR for various proposal schemes.

We find that mAP rises and then falls slightly as the proposal count increases (Fig. 3, solid blue line). This experiment shows that swamping the deep classifier with more proposals does not help, and even slightly hurts, accuracy.

Girshick (2015), https://arxiv.org/abs/1504.08083

Fast R-CNN overview



Fast R-CNN workflow

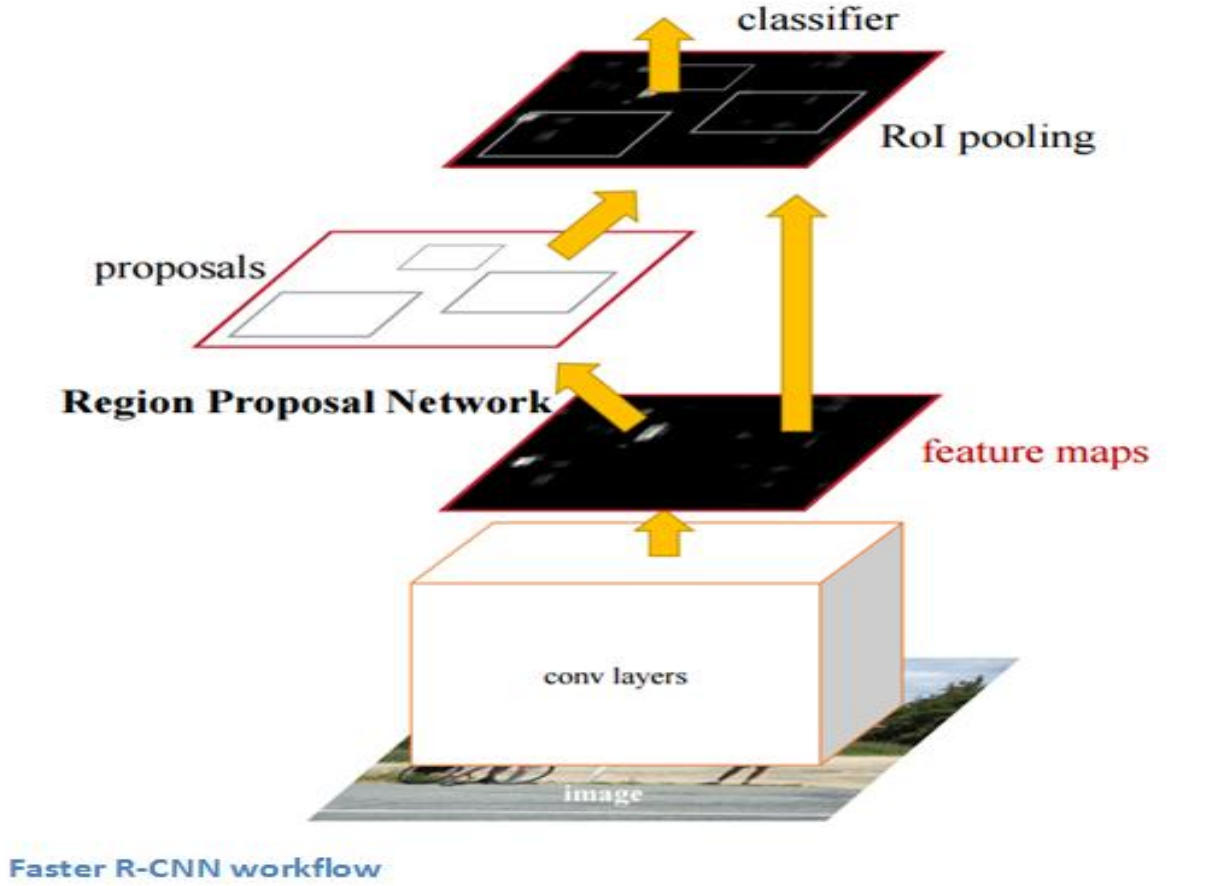
Girshick (2015), https://arxiv.org/abs/1504.08083

Faster R-CNN [1]

- Overcomes bottleneck of region proposal due to selective search
- Region proposal network (RPN) shares convolutional features with detection network
- Uses alternating training (Section 3.2 of paper for further details)
  - 1. Train RPN, and use the proposals to train Fast R-CNN
  - 2. Network tuned by Fast R-CNN is then used to initialize RPN
  - 3. Repeat 1.

[1] Ren et al. (2016), https://arxiv.org/abs/1506.01497

Faster R-CNN overview



Faster R-CNN workflow

Ren et al. (2016), https://arxiv.org/abs/1506.01497

R-CNN comparisons

	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image	50 seconds	2 seconds	0.2 seconds
Speed-up	1x	25x	250x
mAP (VOC 2007)	66.0%	66.9%	66.9%

* Standford lecture notes on CNN by Fei Fei Li and Andrej Karpathy

Mask R-CNN [1]

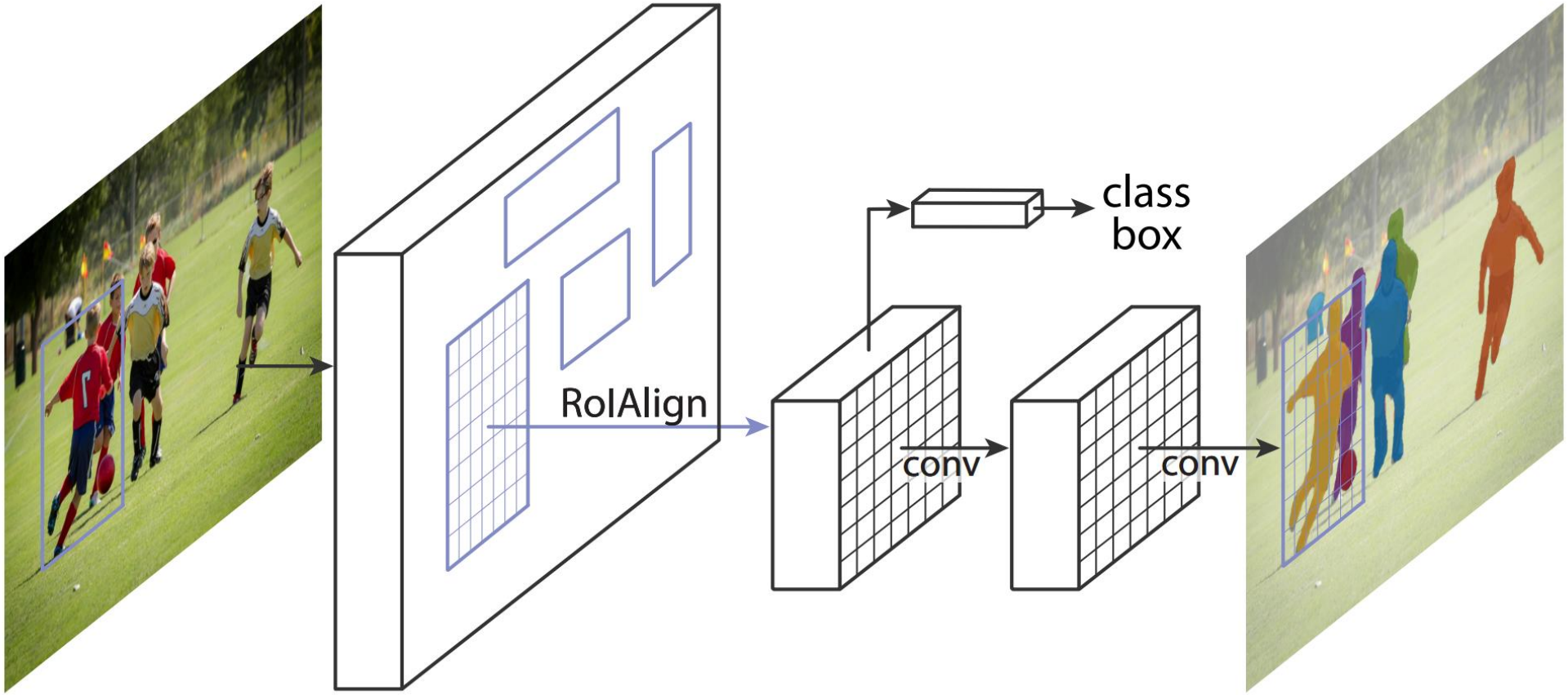
- Instance segmentation
  - Objection detection & Precise segmentation
  - Adds an object mask in parallel to the class label and bounding-box produced by Faster R-CNN for each object

- Segments objects in parallel with detection algorithm
  - Decouples mask and class prediction
  - Fully convolutional network (FCN)
  - Binary mask predicted for each class – without competition between classes

- RoIAlign
  - Use bilinear interpolation to minimize errors due to image re-sizing

[1] He et al. (2017), https://arxiv.org/abs/1703.06870

Mask R-CNN overview



He et al. (2017), https://arxiv.org/abs/1703.06870

Figure 4. More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).