Mask RCNN

Presented by: Muhammad Tayyab

Overview

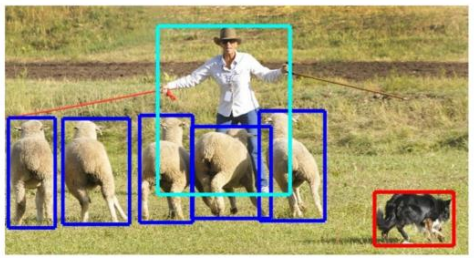
- Background
  - RCNN (CVPR 14)
  - FastRCNN (ICCV 15)
  - FasterRCNN (NIPS 15)
  - MaskRCNN (ICCV 17)
- Network Backbone
  - Region Proposal Network
  - RoI Pooling
  - RoI Align
- Network Head
  - Classification & Detection
  - Segmentation
- Experiments

Background

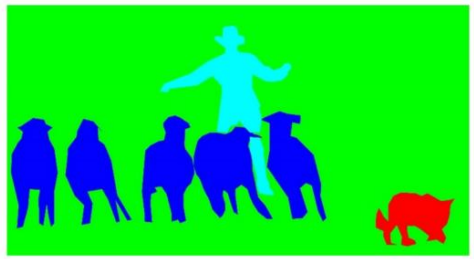
- State of the art multitask model for visual scene understanding
  - object detection
  - Classification
  - instance segmentation



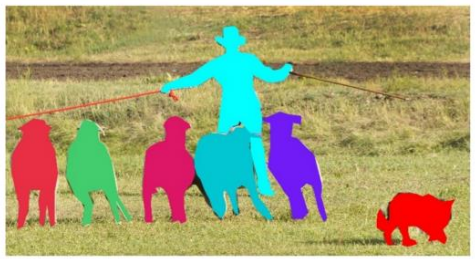
Image classification



Object detection

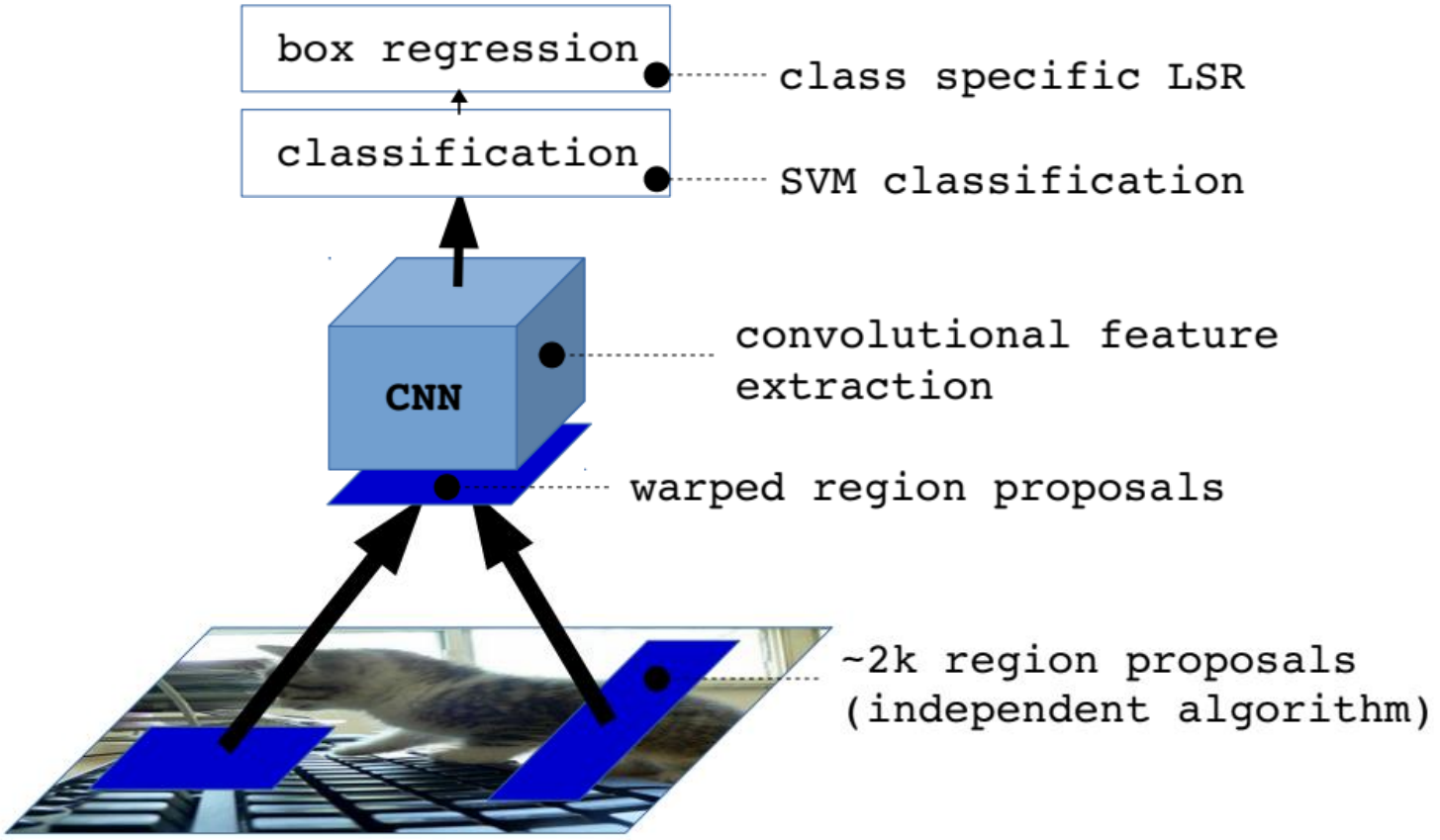


Semantic segmentation



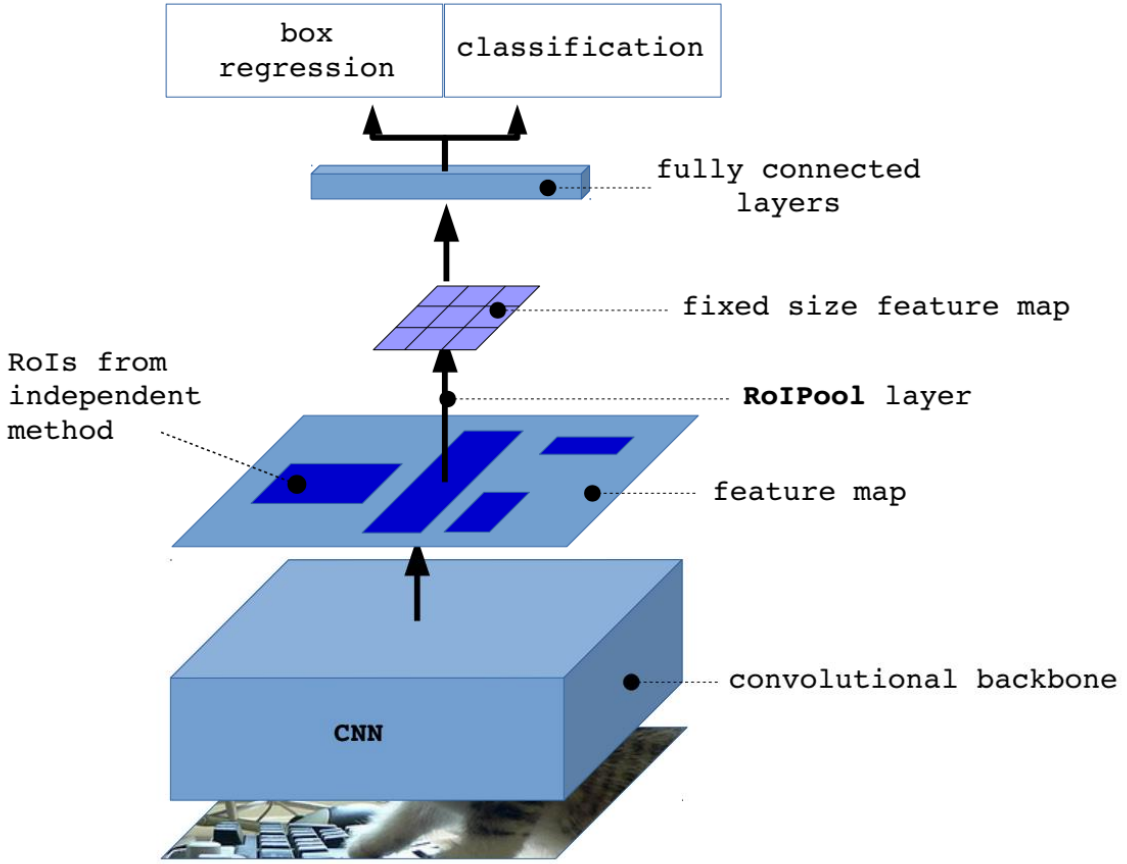
Instance segmentation

Background-RCNN

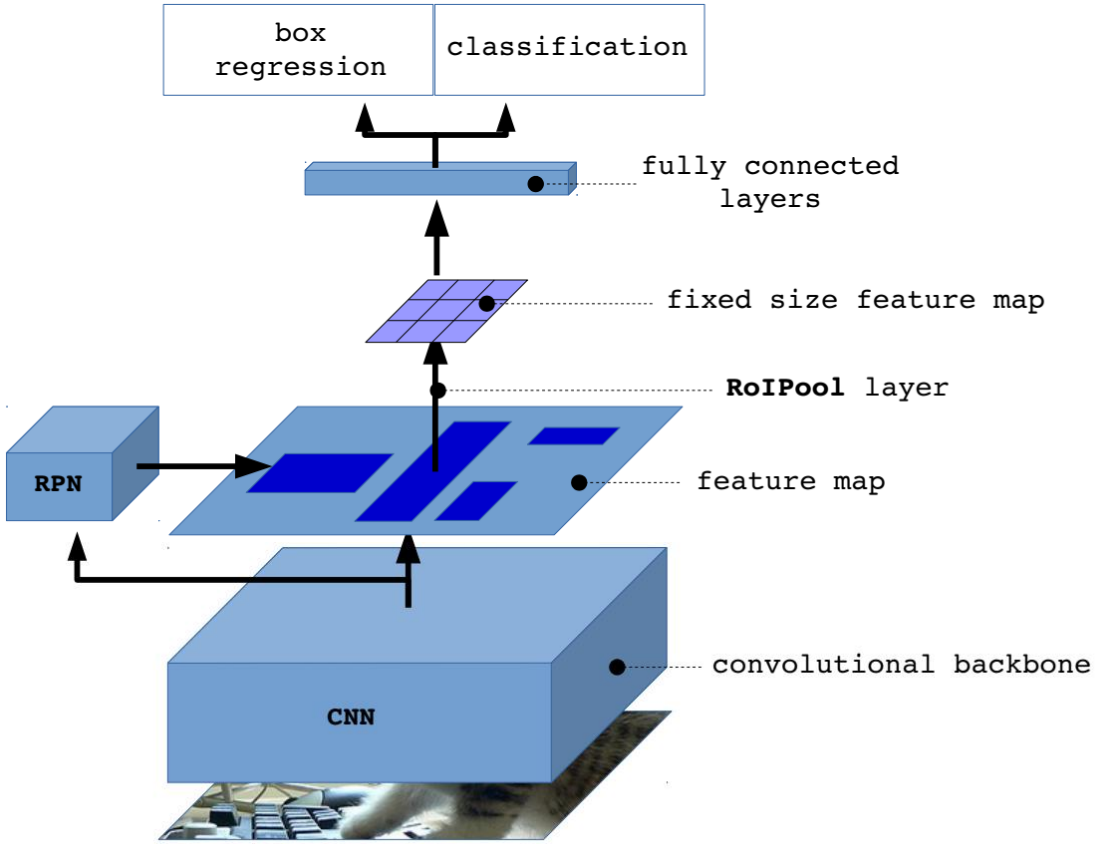


Source: https://lmb.informatik.uni-freiburg.de/lectures/seminar_brox/seminar_ss17/maskrcnn_slides.pdf

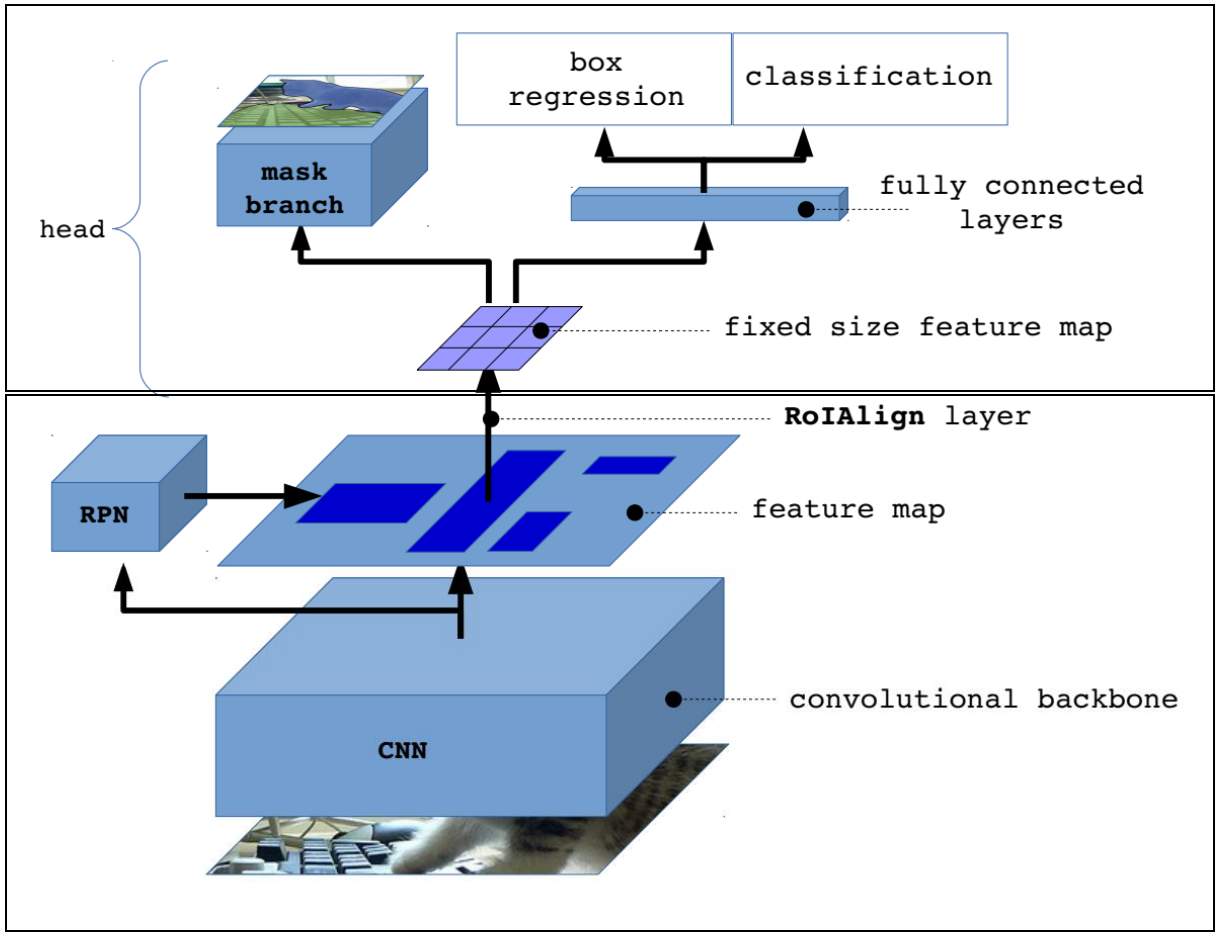
Background-FastRCNN



Background-FasterRCNN



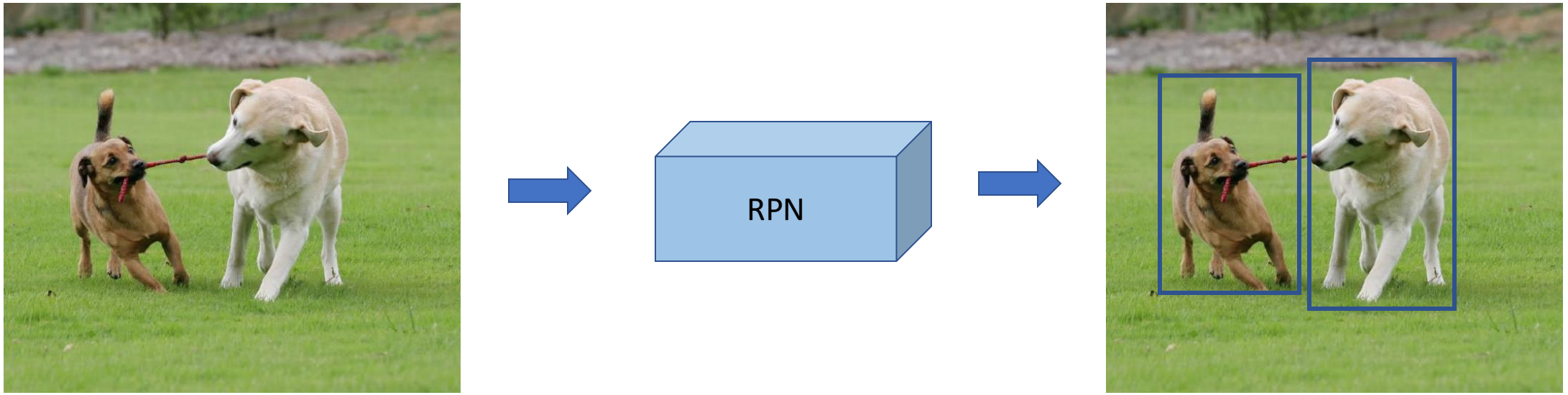
MaskRCNN



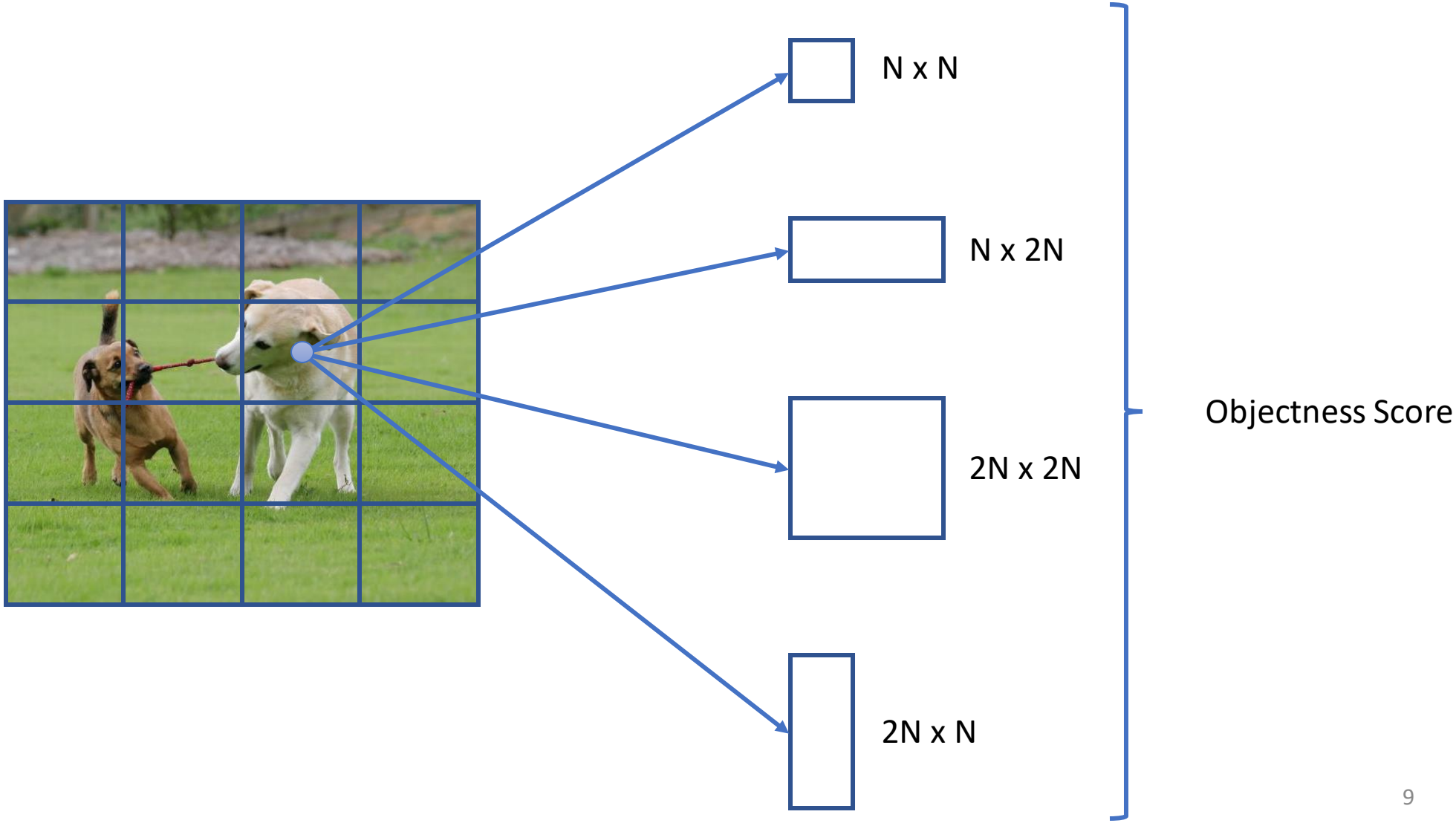
Head

Backbone
- Resnet50
- Resnet100
- FPN

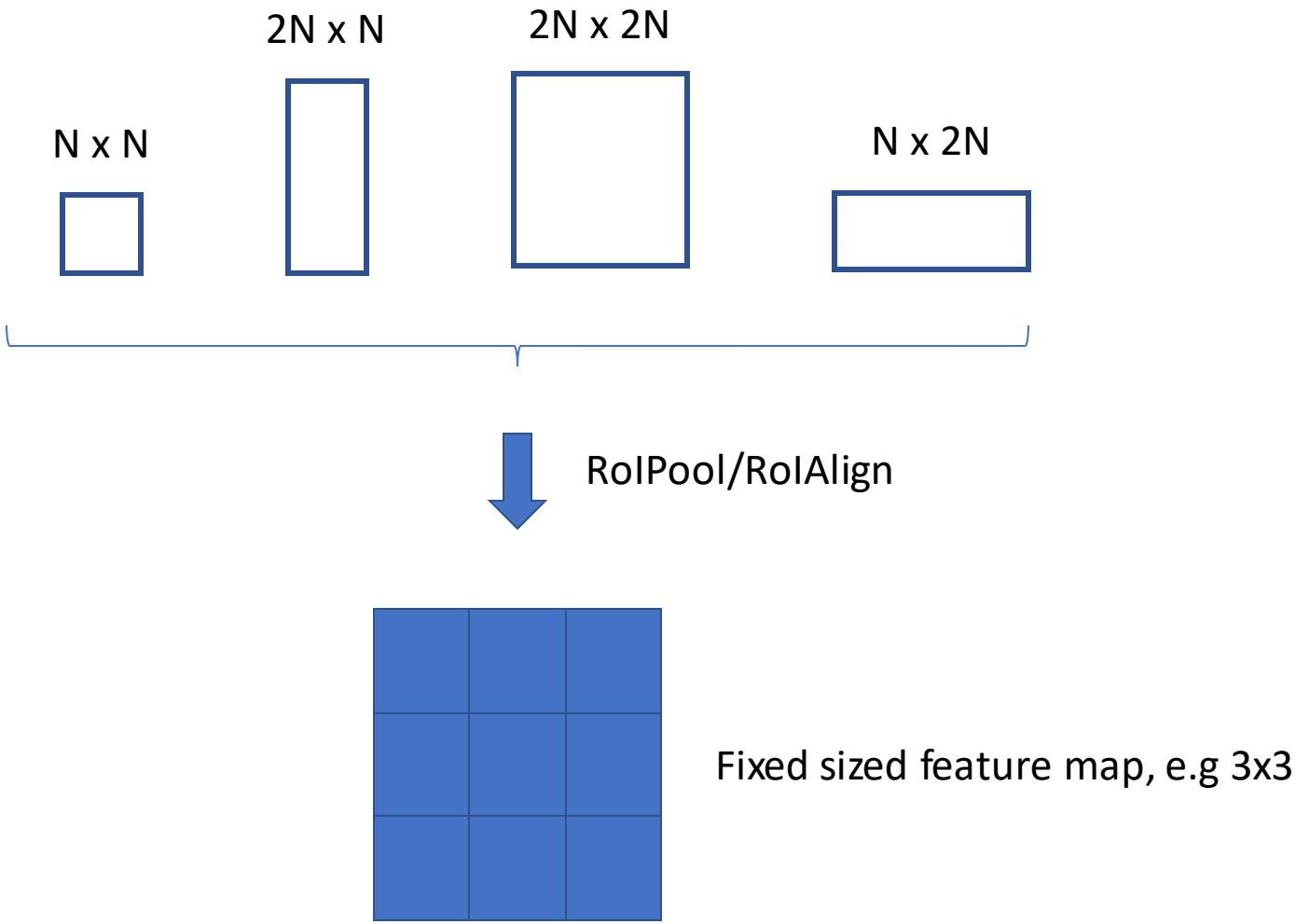
Region Proposal Network



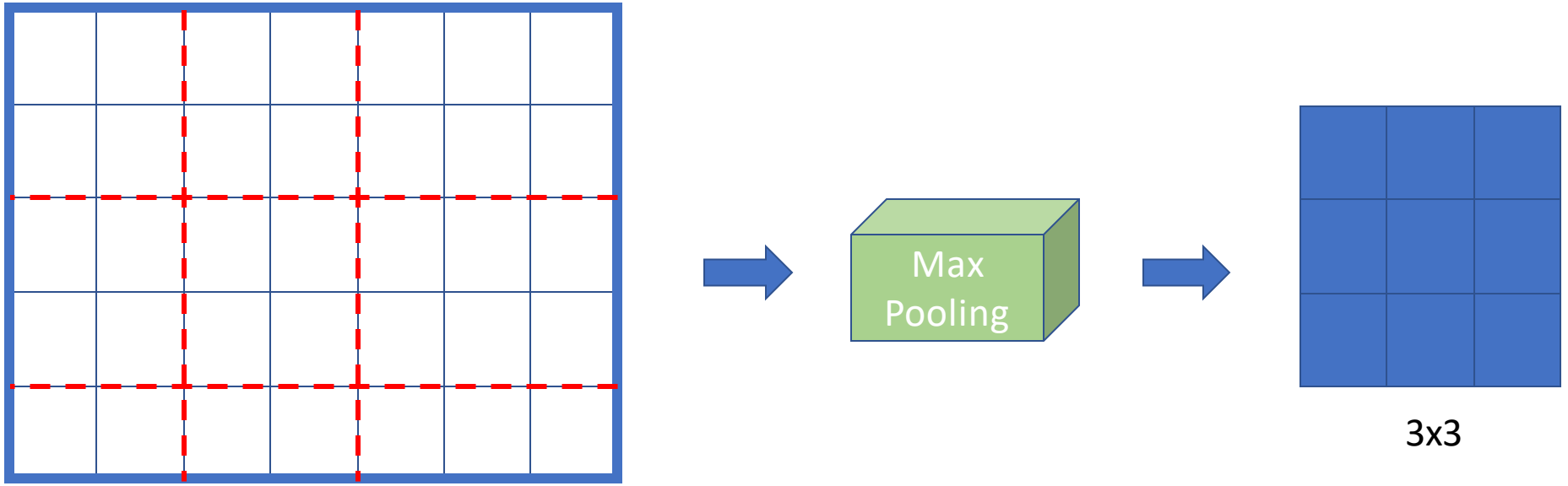
Region Proposal Network



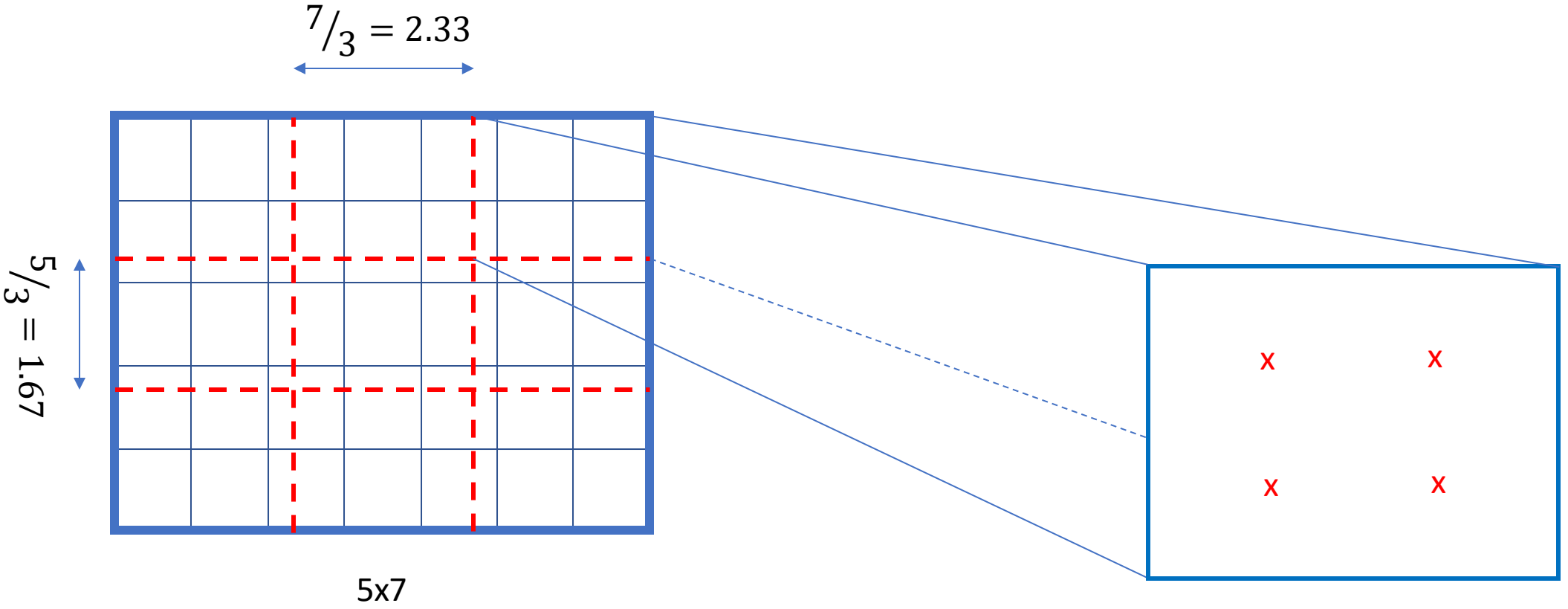
RoI feature extraction



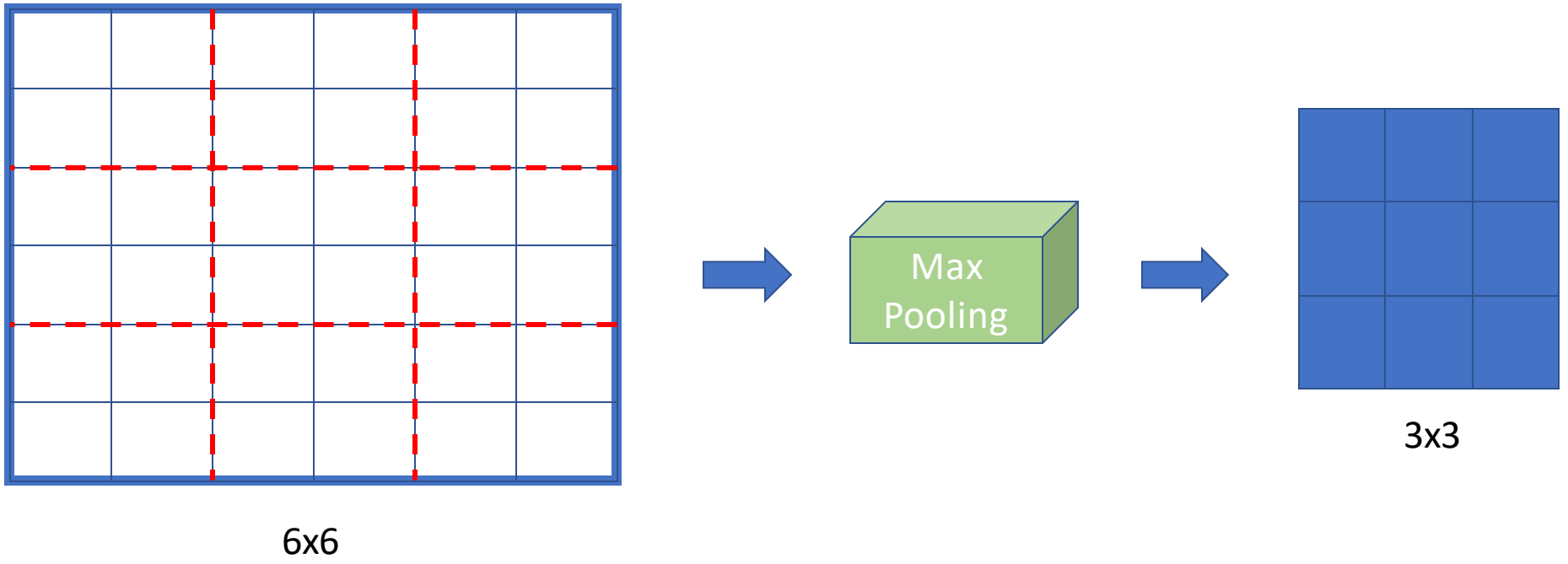
RoI Pooling-FastRCNN



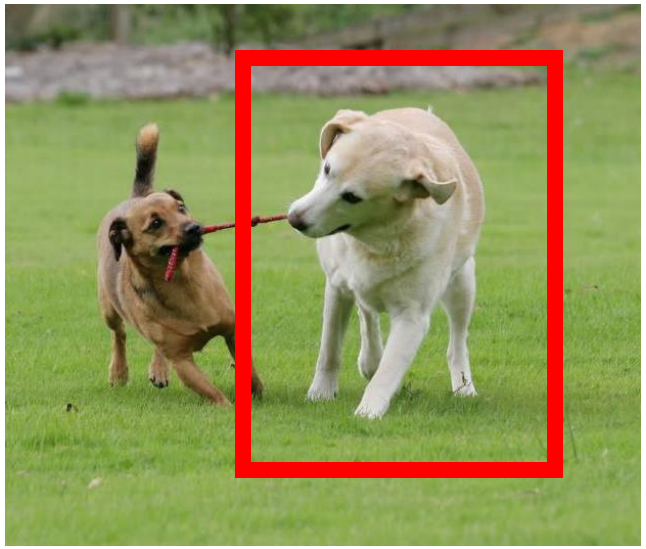
RoI Align-MaskRCNN

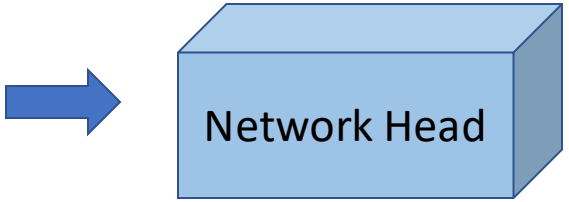


RoI Align-MaskRCNN



Head





Dog





Head – Classification & Detection

- Fully connected branch:
  - K + 1 softmax for classification
  - 4K box regression targets
    - (x, y, w, h)
- Multitask loss:
  - L_{cls}: negative log likelihood
  - L_{reg}: smooth L_1 loss

$$L_{box} = L_{cls} + L_{reg}$$

Head – Mask Prediction

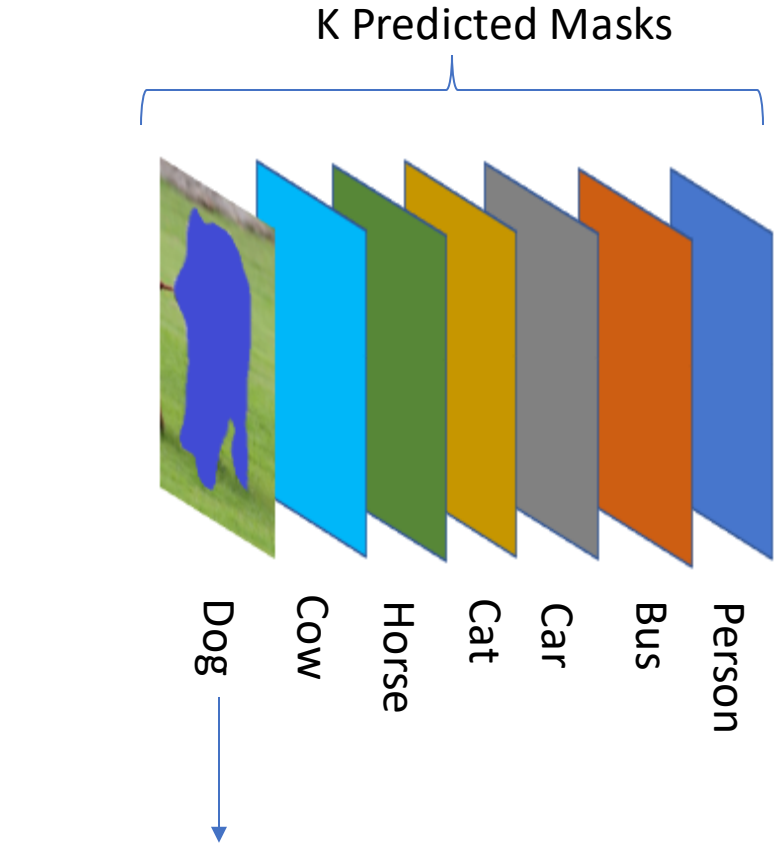
- Fully convolutional
- K · (m × m) sigmoid outputs
  - pixel-wise binary classification
  - one mask for each class
- L_{mask}: mean binary cross-entropy

$$L_{box} = L_{cls} + L_{reg}$$

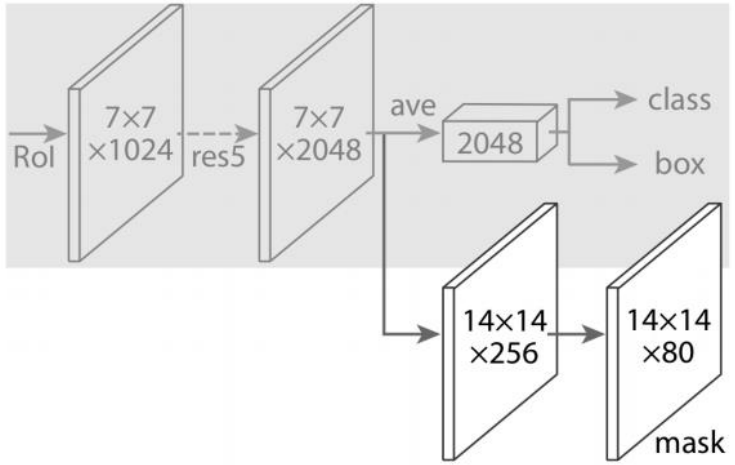
$$L_{final} = L_{box} + L_{mask}$$

Mask Loss

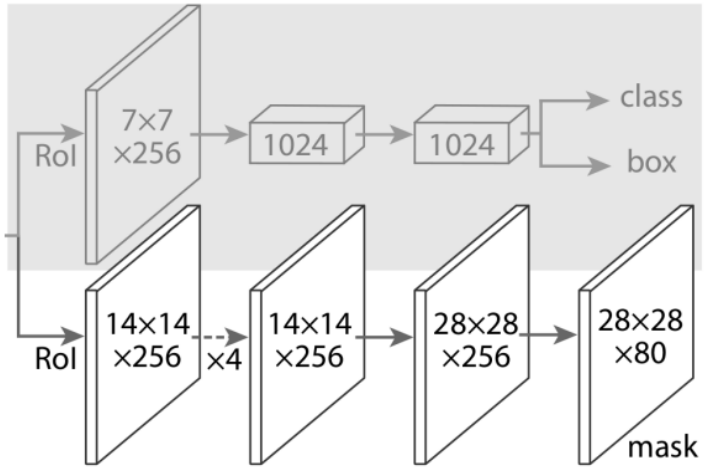



K Predicted Masks
Dog
Cow
Horse
Cat
Car
Bus
Person
Used to compute loss
fo mask Prediction

Head



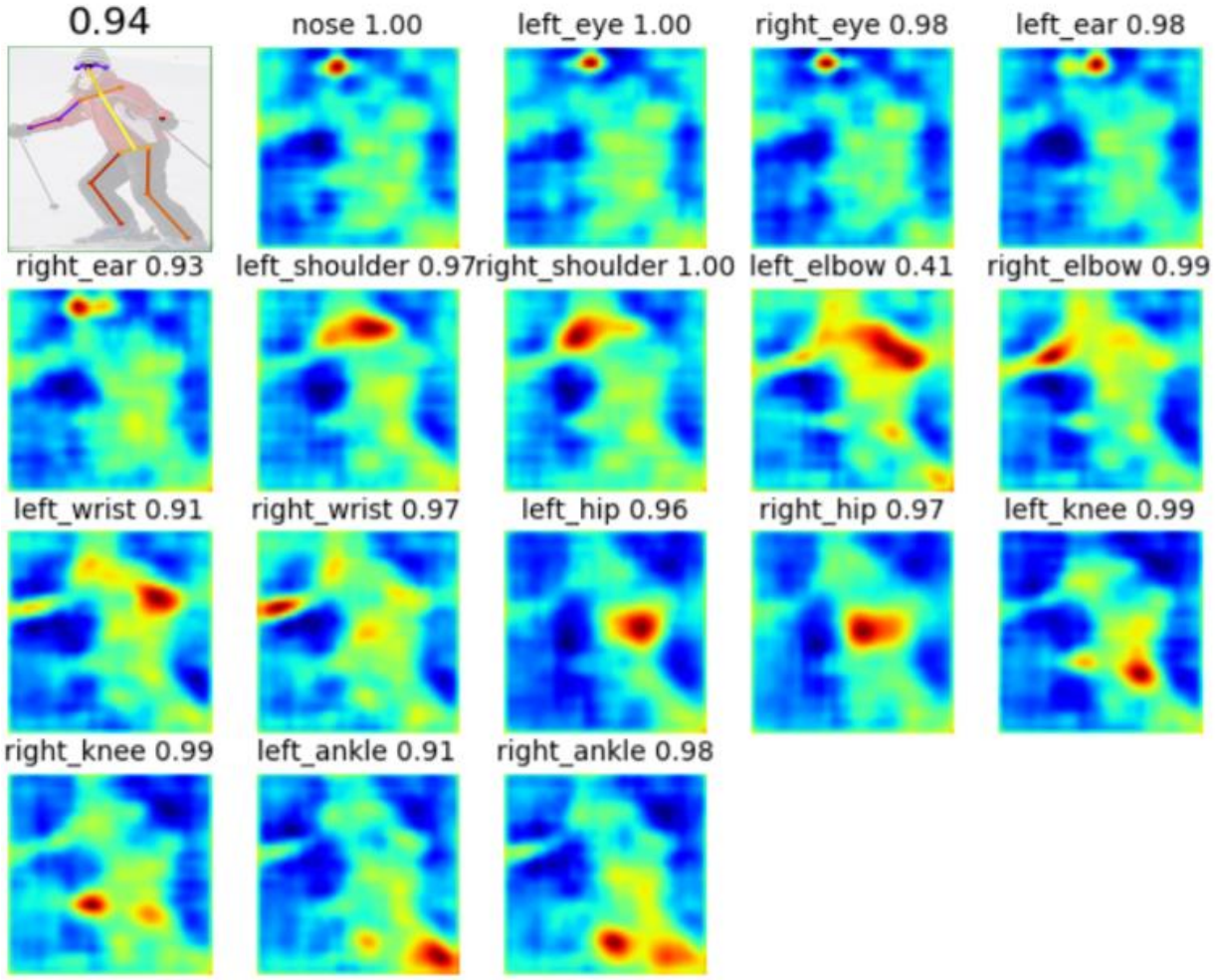
Extended Faster R-CNN head, on ResNet-C4 feature map



Extended Faster R-CNN head, on FPN feature map

Human Pose Estimation

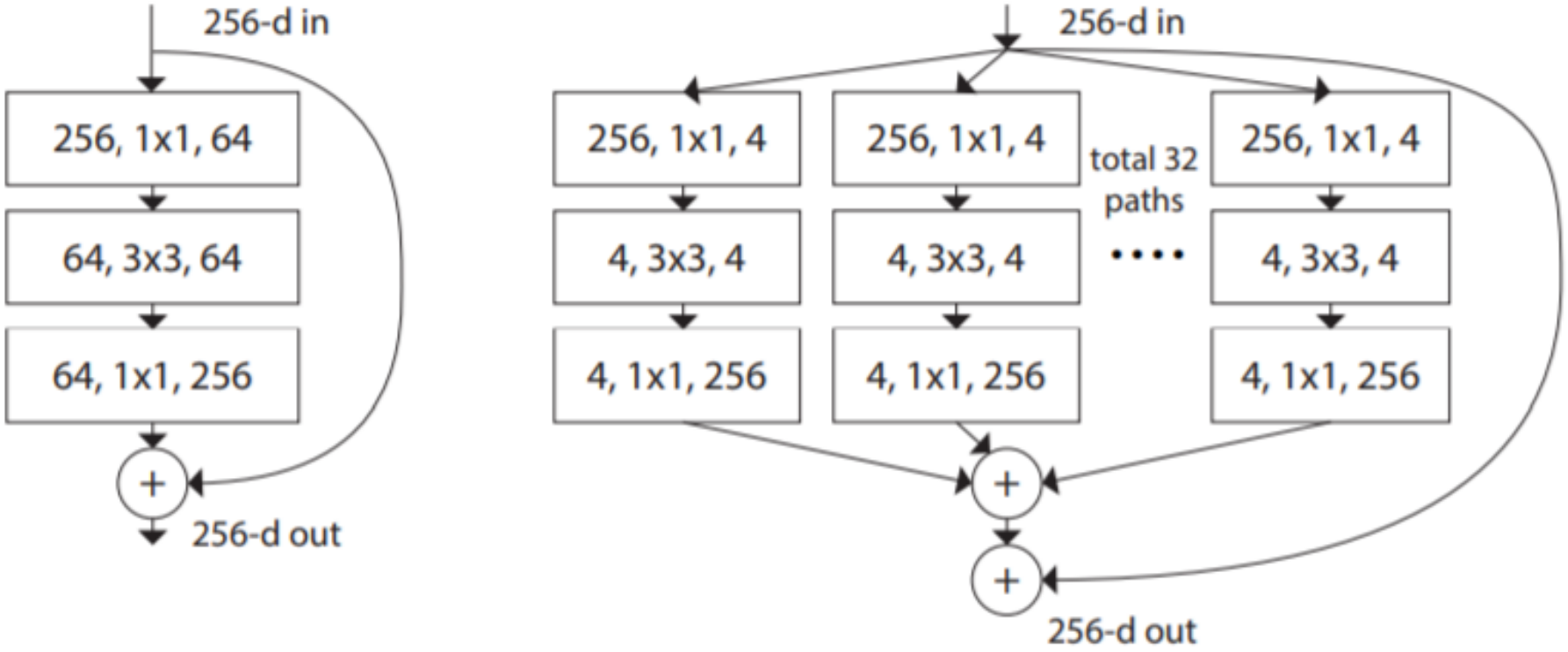
- 1 keypoint = 1-hot “mask”
- Human pose = 17 masks



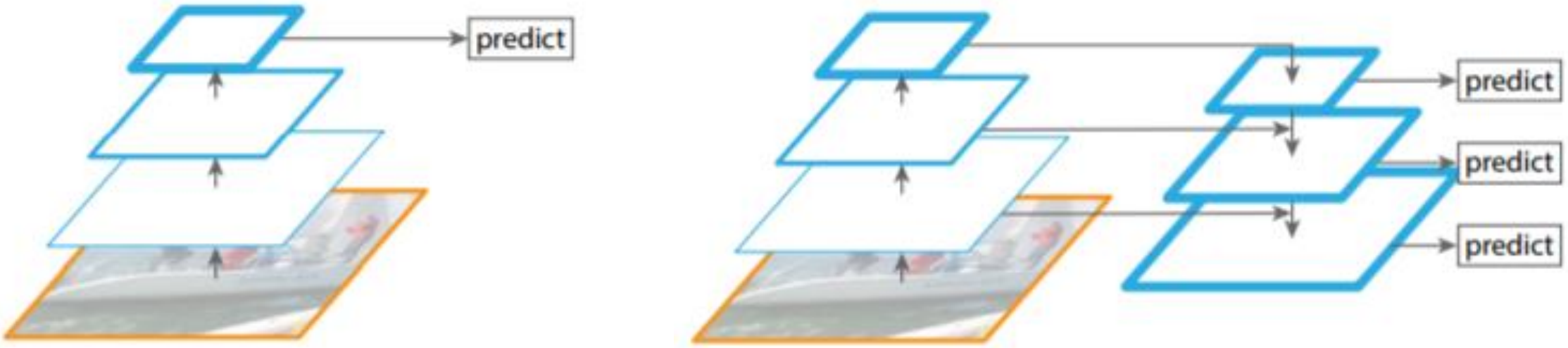
Source: http://kaiminghe.com/iccv17tutorial/maskrcnn_iccv2017_tutorial_kaiminghe.pdf

Experiments

Resnet VS ResNeXt



Single Feature Map VS FPN



Datasets and Metrics

- Main dataset: MS COCO
  - 80 classes
  - 115k training images
- Similarity measure: Intersection over Union (IoU)
  - AP_{50} & AP_{75} (PASCAL VOC metrics):
    - Average Precision: IoU threshold (.50, .75)
  - AP (MS COCO metric): mean Average Precision over different IoU thresholds.

Instance Segmentation

Model	backbone	AP	AP_{50}	AP_{75}
MNC	ResNet-101-C4	24.6	44.3	24.8
FCIS+++	ResNet-101-C5-dil.	33.6	54.5	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**

On COCO Dataset

Object Detection

Model	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
Faster R-CNN	ResNet-101-FPN	36.2	59.1	39.0
Faster R-CNN	same + RoIAlign	37.3	59.6	40.3
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**

On COCO Dataset

Ablation

Backbone architecture:

net-depth-features	AP	AP_{50}	AP_{75}
ResNet-50-C4	30.3	51.2	31.5
ResNet-101-C4	32.7	54.2	34.3
ResNet-50-FPN	33.6	55.2	35.3
ResNet-101-FPN	35.4	57.3	37.5
ResNeXt-101-FPN	**36.7**	**59.5**	**38.9**

Ablation

RoIAlign layer:

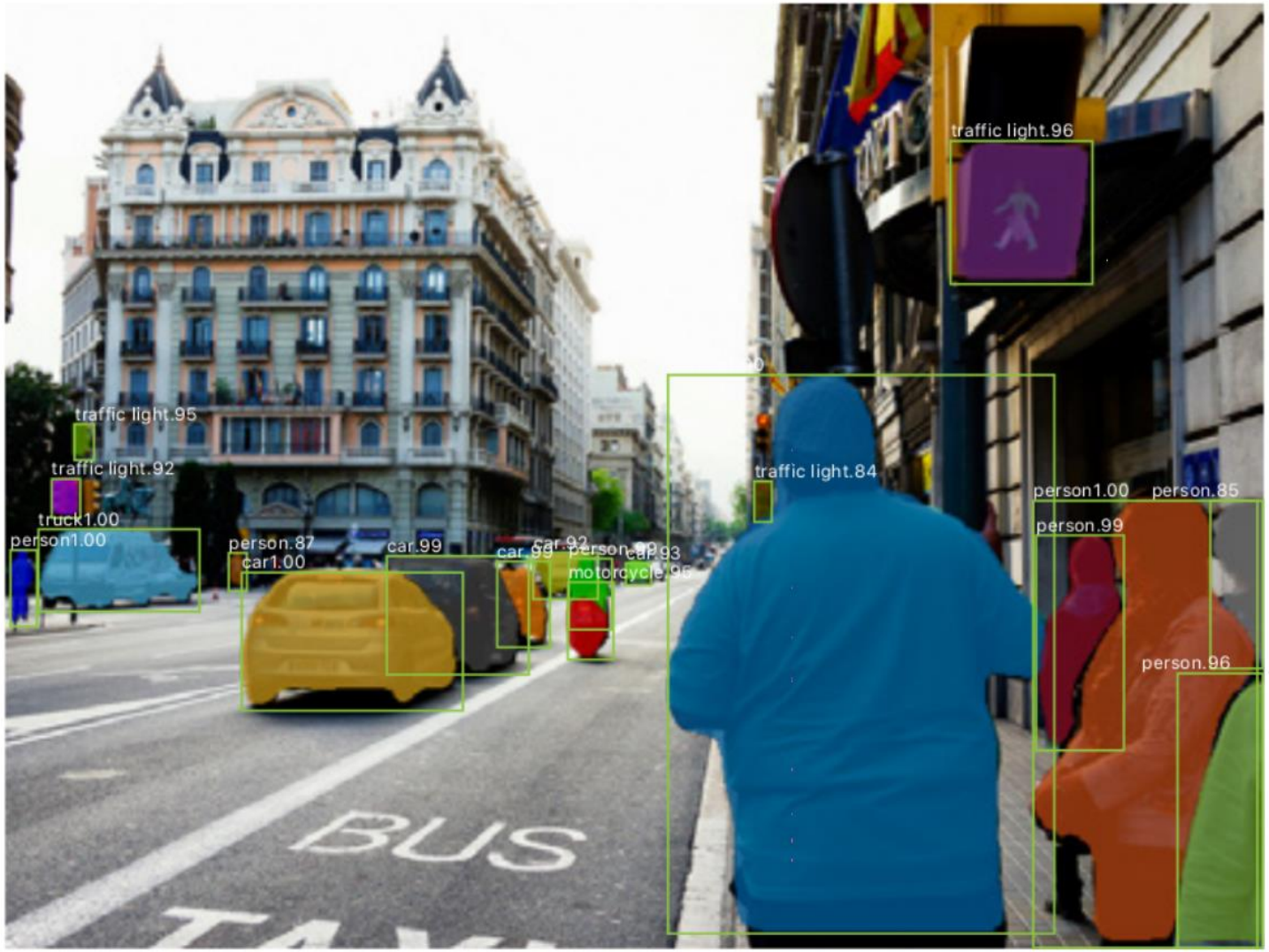
	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
RoIPool	23.6	46.5	21.6	28.2	52.7	26.9
RoIAlign	**30.9**	**51.8**	**32.1**	**34.0**	**55.3**	**36.4**
	+7.3	*+ 5.3*	*+10.5*	*+5.8*	*+2.6*	*+9.5*

Ablation

Mask branch independence:

last fc layer	AP	AP_{50}	AP_{75}
multinomial loss	24.8	44.1	25.1
binomial loss	**30.3**	**51.2**	**31.5**
	+5.5	+7.1	+6.4

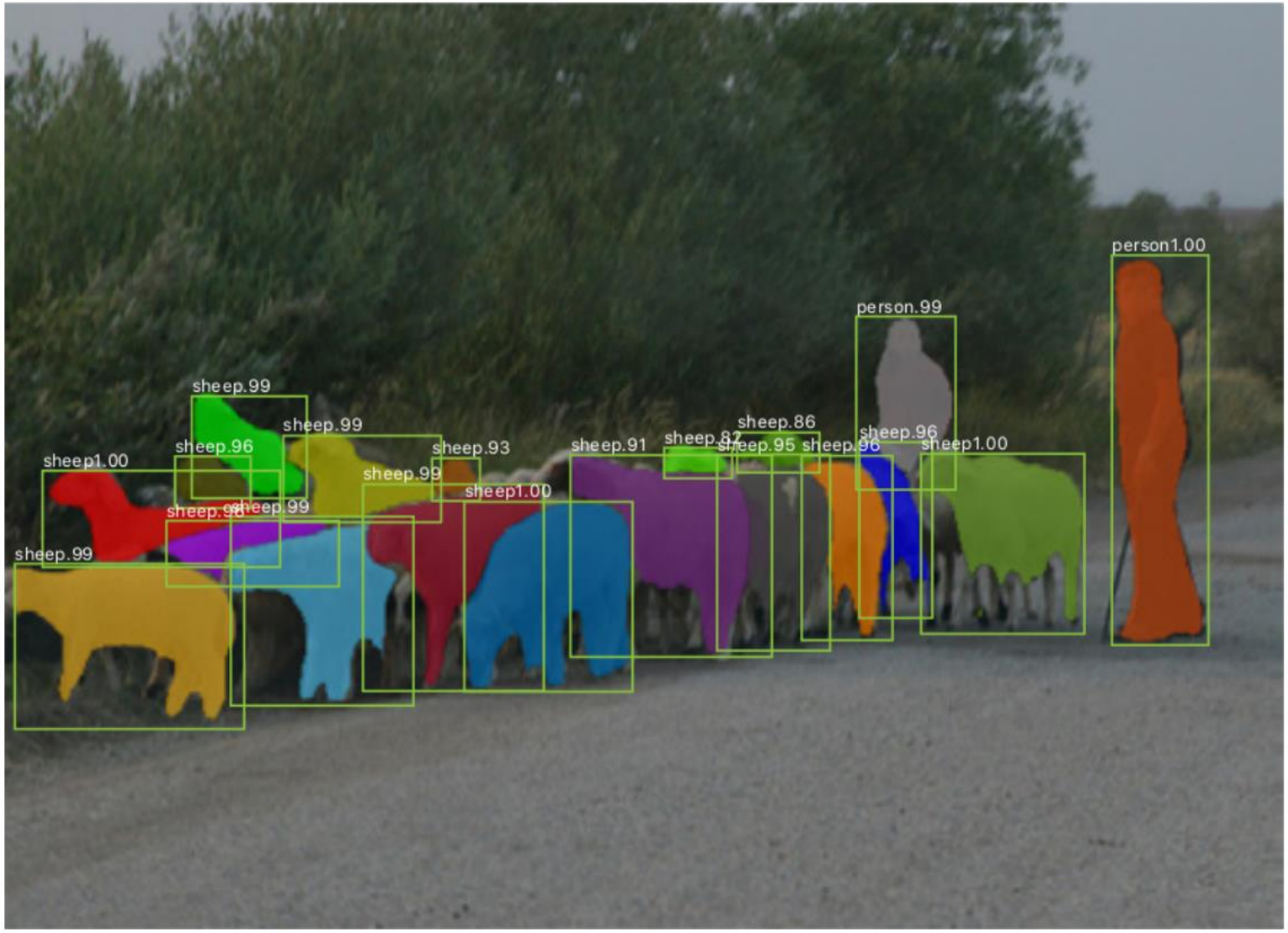
Visual Results



Visual Results



Visual Results



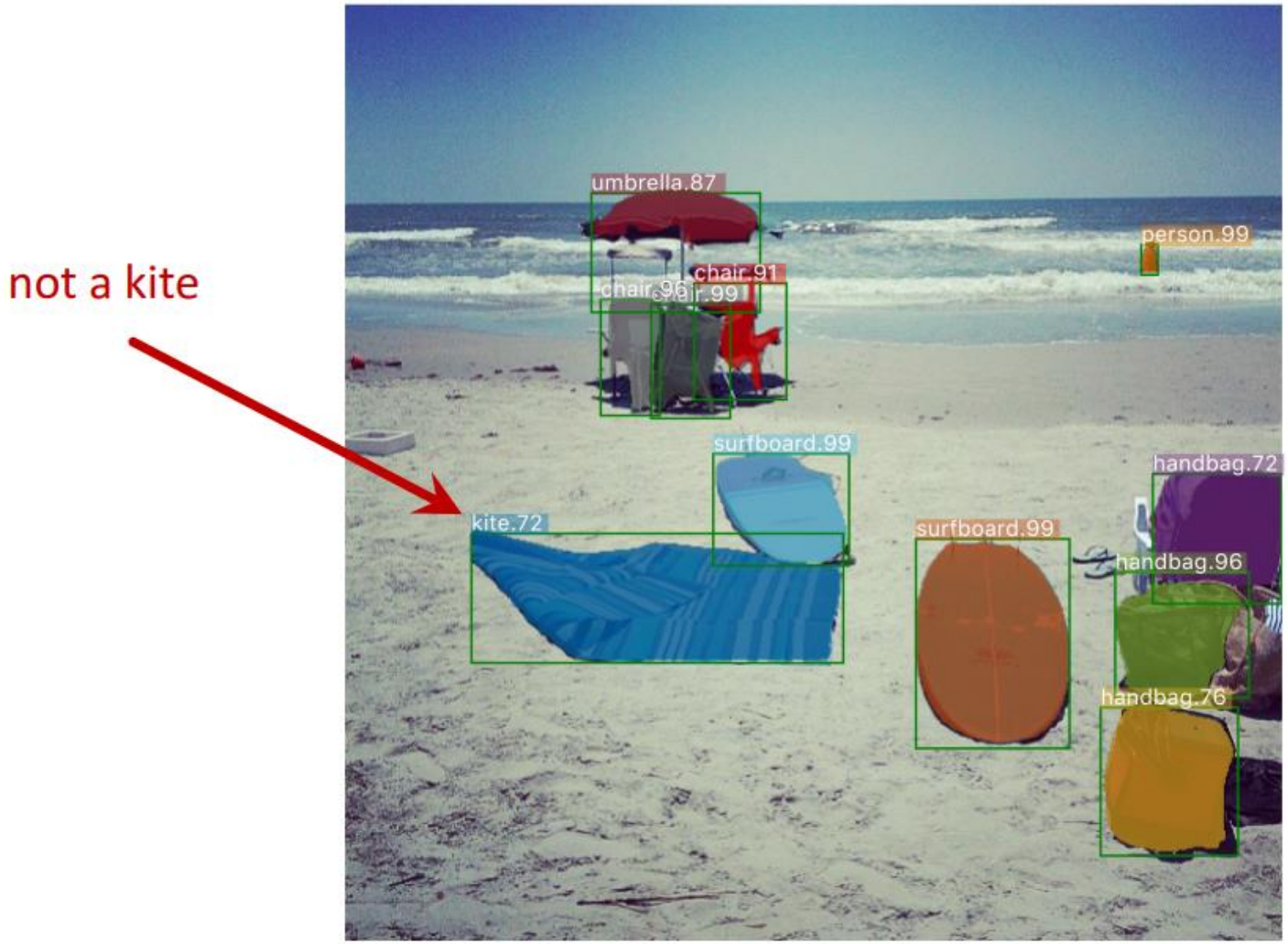
Visual Results



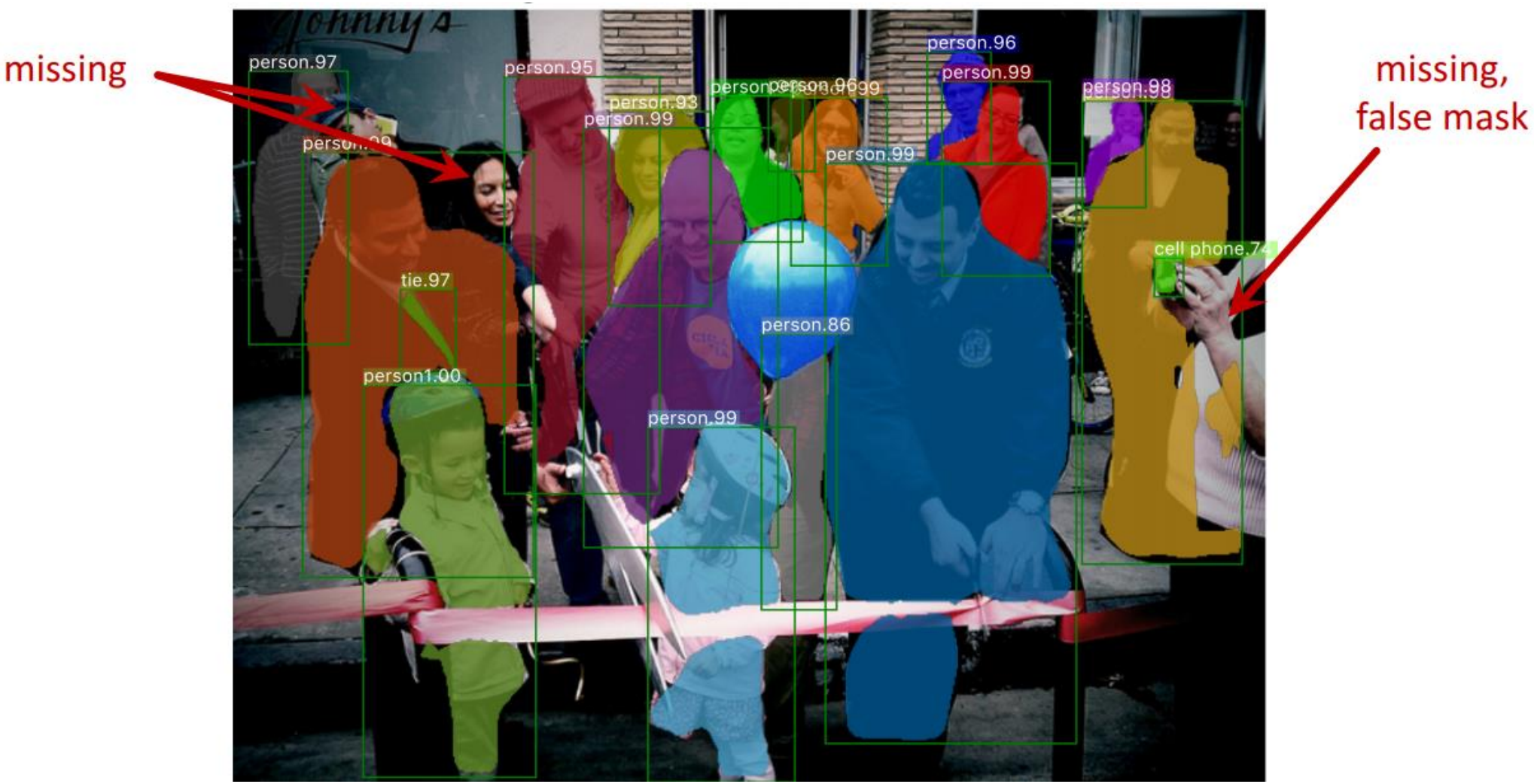
Visual Results



Failure Cases



Failure Cases



Thank You