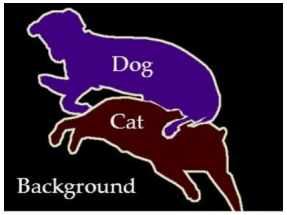


Mask R-CNN

Kaiming He, Georgia, Gkioxari, Piotr Dollar, Ross Girshick

Presenters: Xiaokang Wang, Mengyao Shi

Feb. 13, 2018

Common computer vision tasks

Image Classification: one label is generated for an image from the probability from a softmax function.

There is a cat in the image.

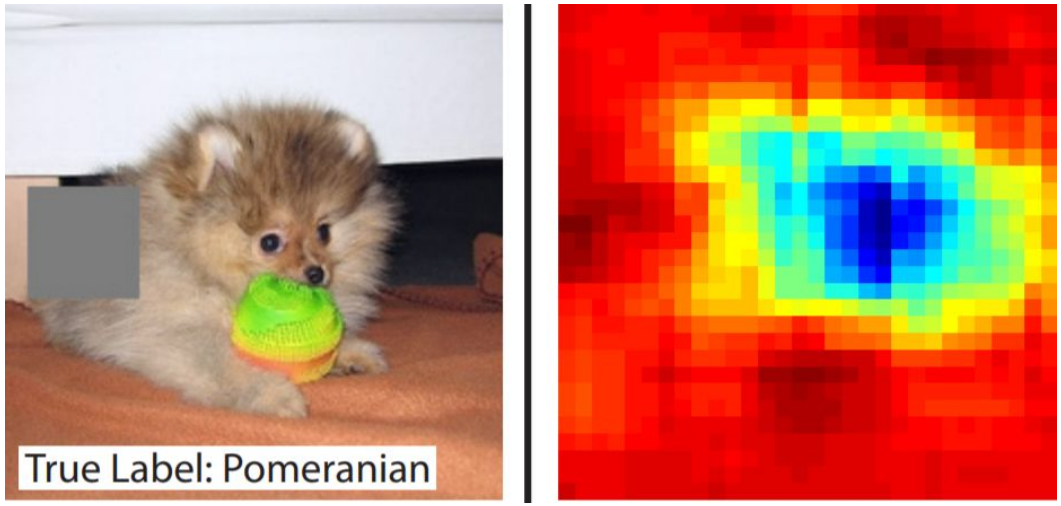
Common computer vision tasks

Image Classification: one label is generated for an image from the probability from a softmax function.

There is a cat in the image. **Where is the cat?**

Common computer vision tasks

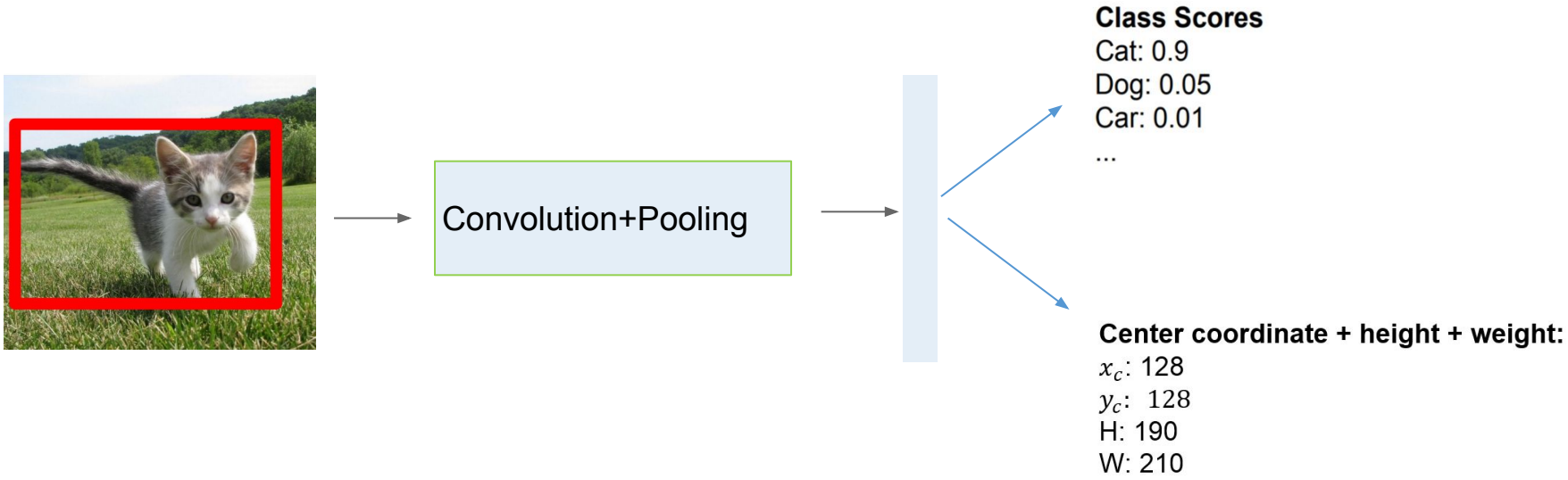
A trained neural network roughly knows where the object is.



Ref: Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

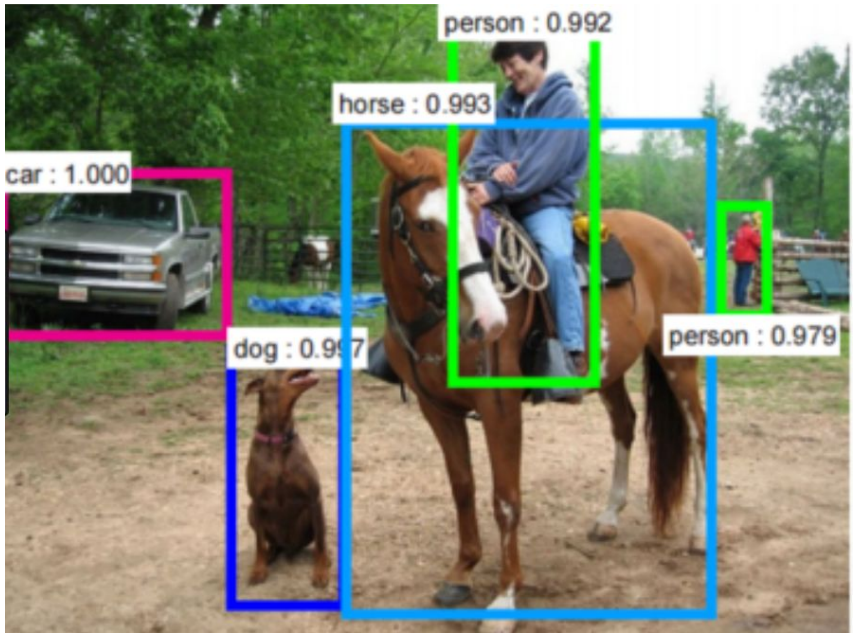
Common computer vision tasks

Classification+Localization: one label is generated for an image based on the probability from a softmax function.



Common computer vision tasks

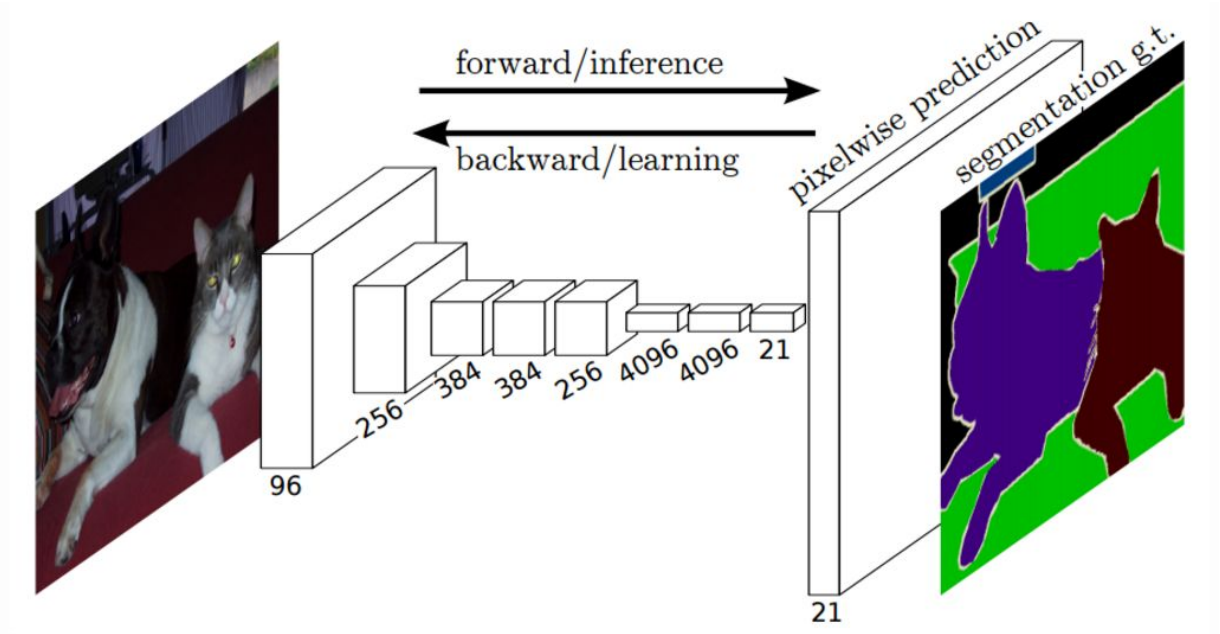
Object detection: detect the presence of objects and then **label** and **localize** them.



Semantic segmentation

Upsampling the features to **the same witdth and height** as the input image.

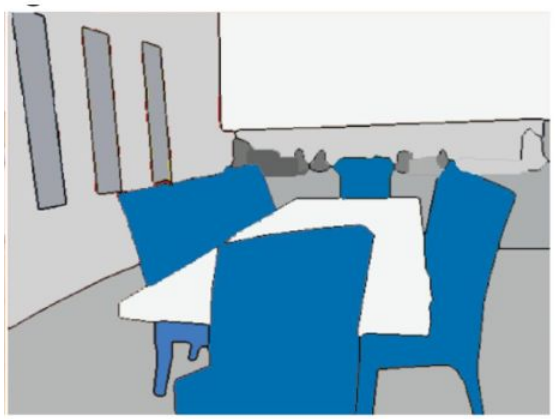
1. **No fully connected layers** in such an architecture
2. **1×1 convolution** is used to collapse all the channels of a feature volume into one chanel



Ref: Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "**U-net:** Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

Sematic segmentation vs instance segmentation

Input Image



Semantic Segmentation

Semantic Instance Segmentation

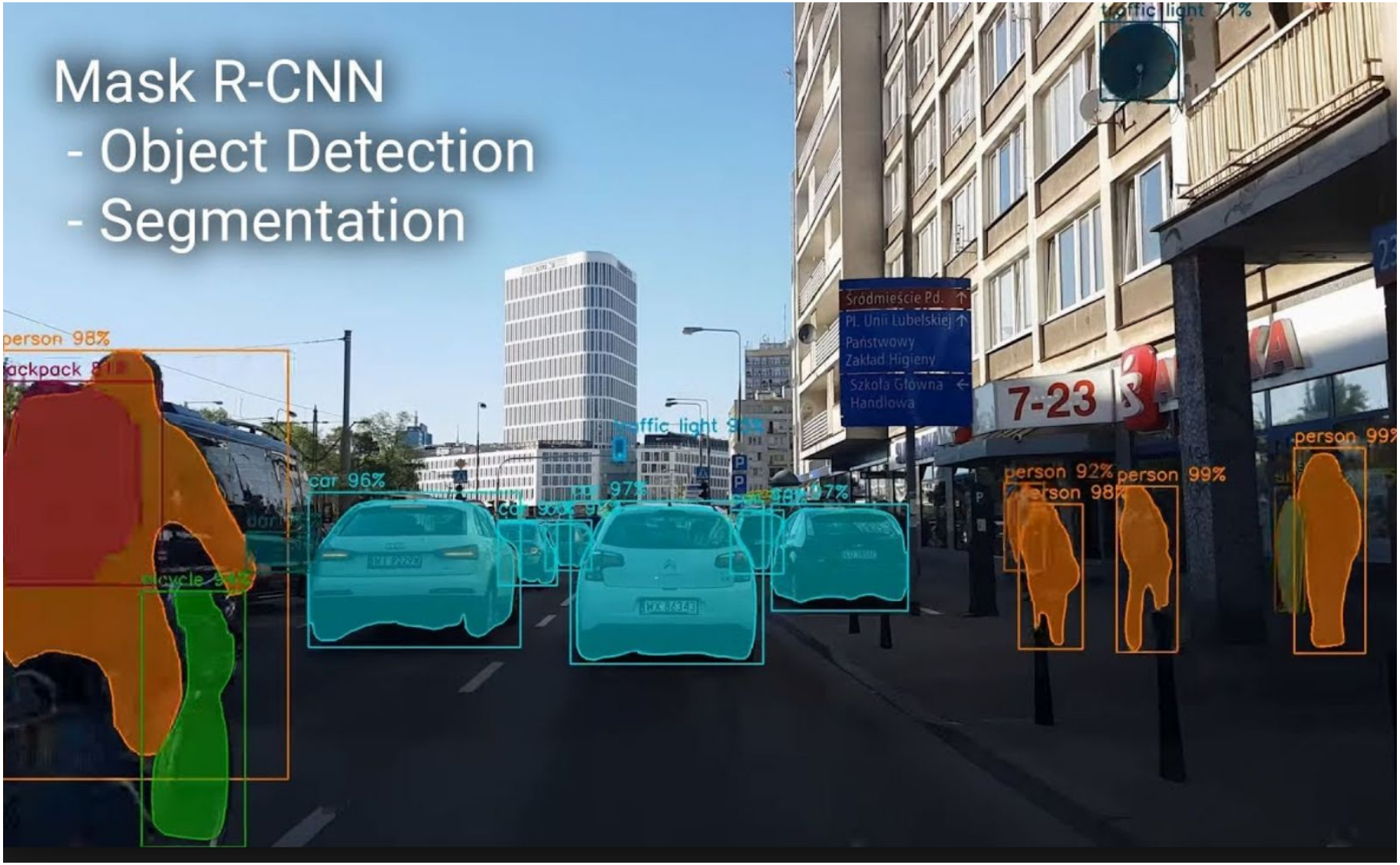
Source;
https://stackoverflow.com/questions/33947823/what-is-semantic-segmentation-compared-to-segmentation-and-scene-labeling

Sematic segmentation vs instance segmentation

One pixel can have **different semantic meaning** in different regions

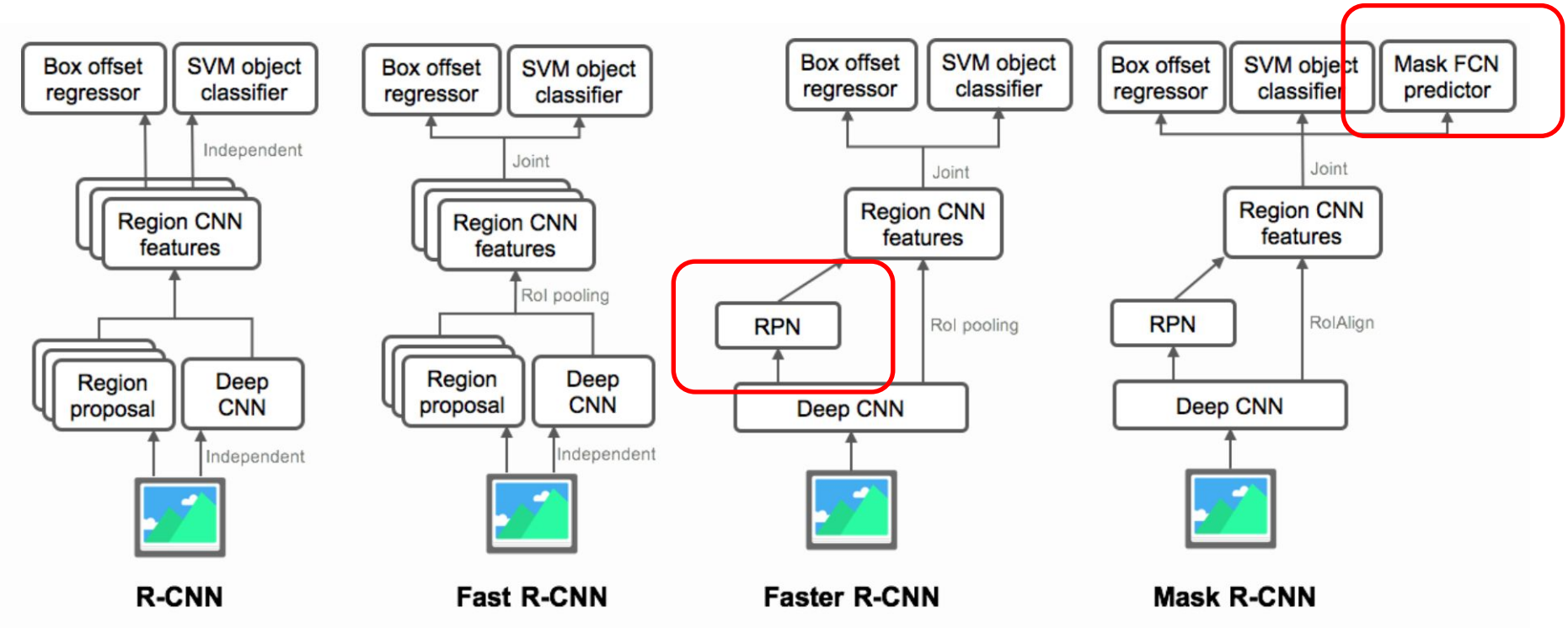
Ref: Yi Li, et al. FCIS

What can Mask-RCNN do?



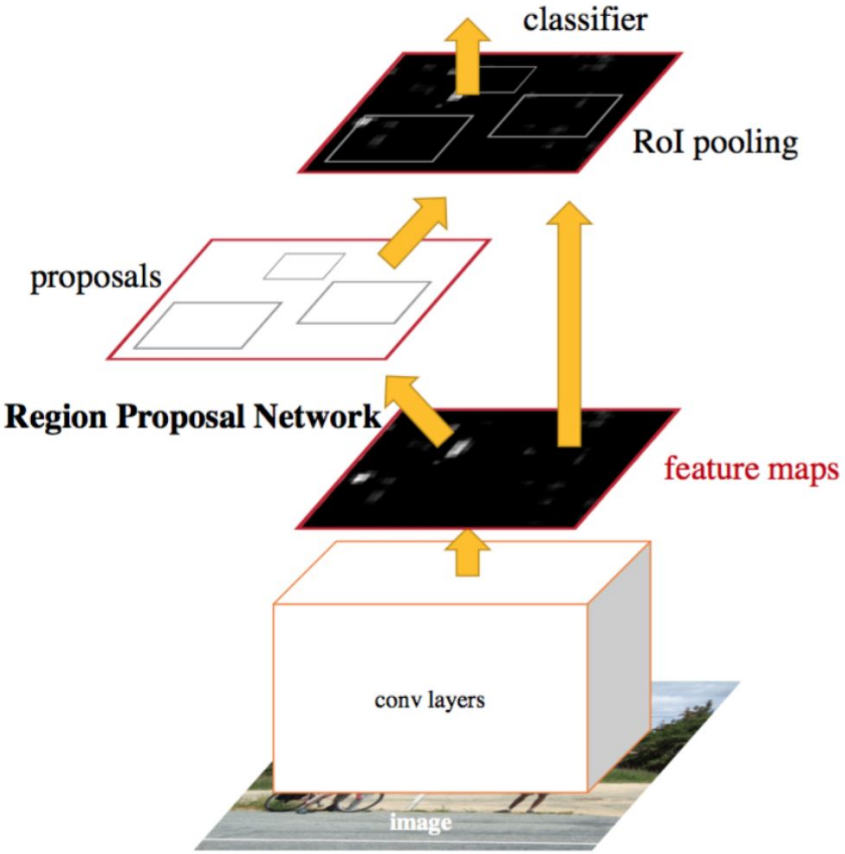
How does Mask-rcnn work at a high level?

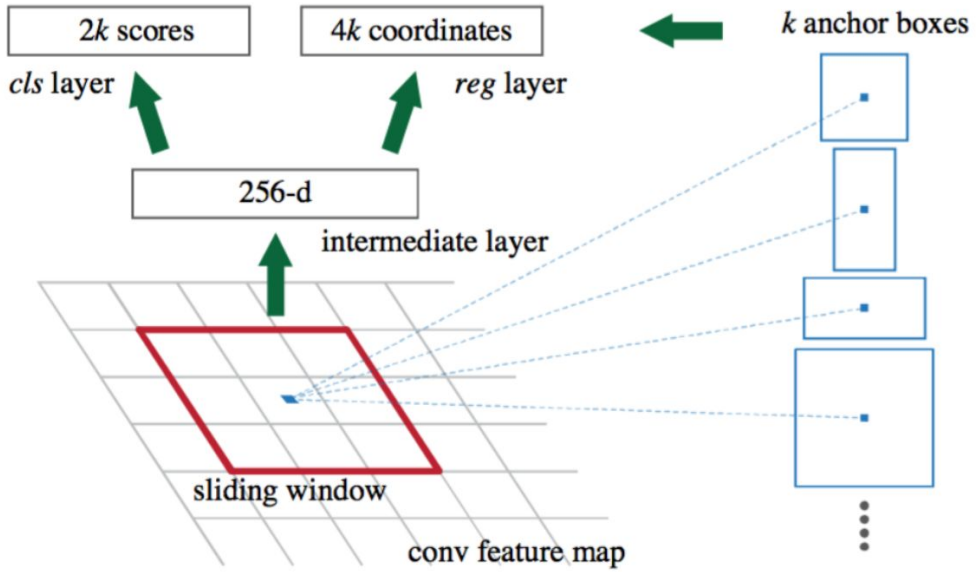
An **FCN** on RoIs is added to **Faster-rcnn**



source: https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html

Faster-rcnn = Fast-rcnn + RPN

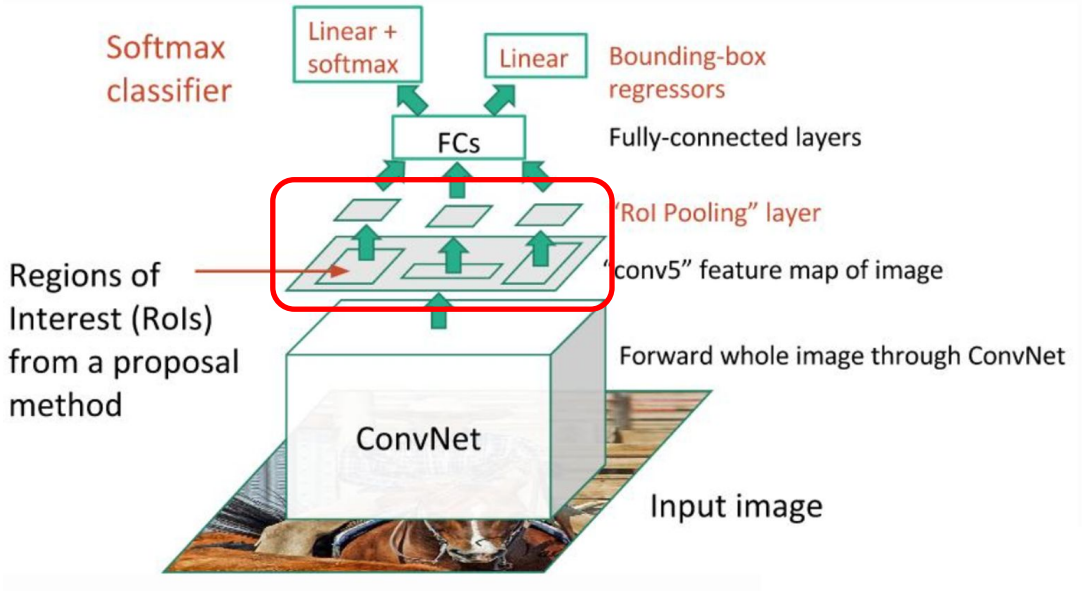


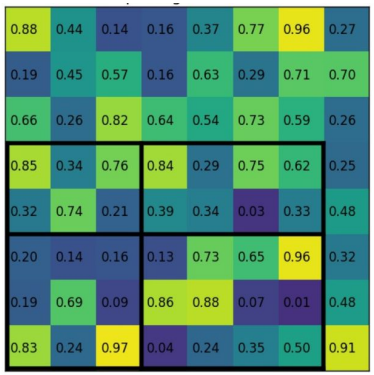


Ref: Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks."

RoIPool layer in fast-rcnn

RoI pooling layer uses max pooling to **covert** the features inside **any valid region** of interest **into a small feature map with a predefined size.**

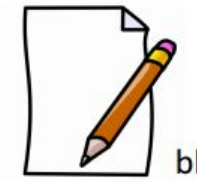




RoI pooling

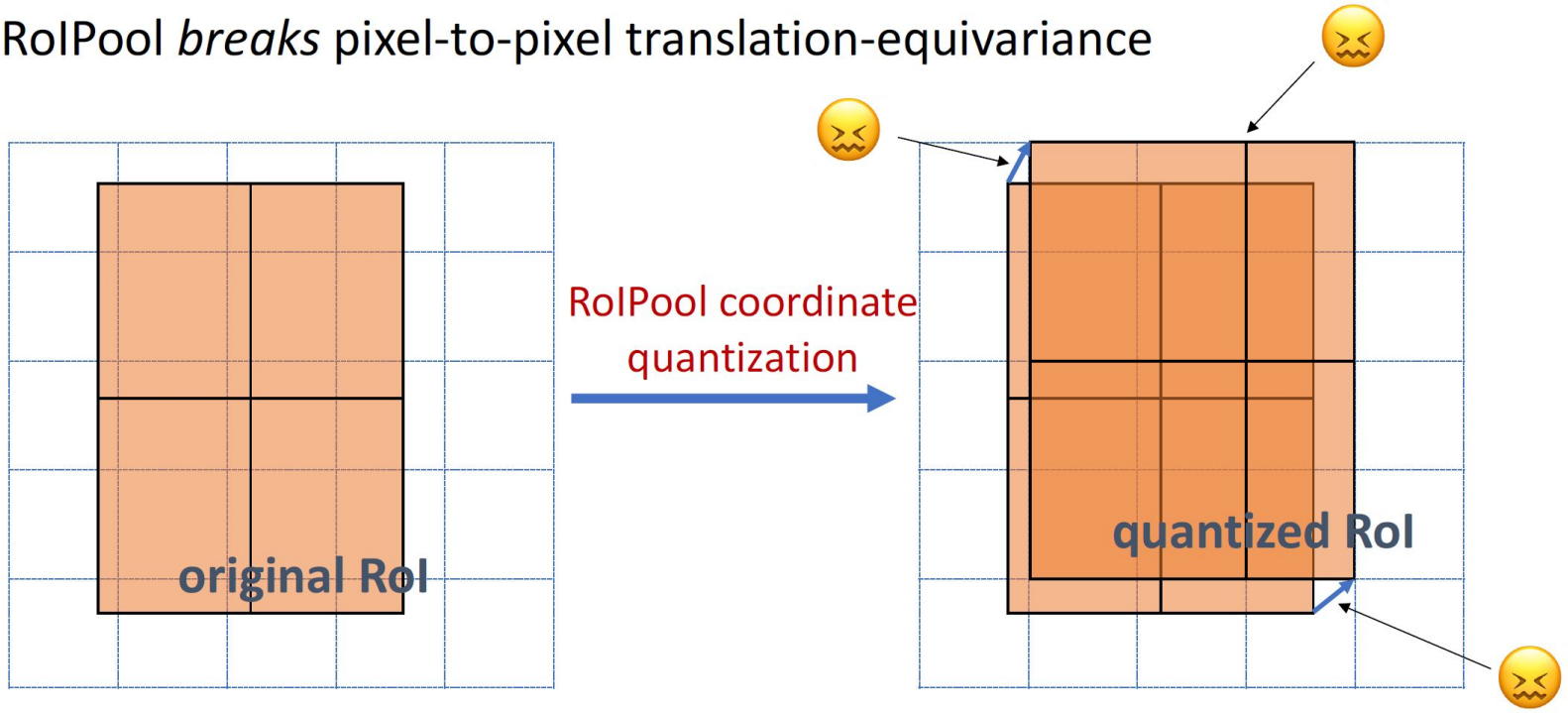
Divide the h×w **RoI** window **into a H×W grid** of subwindows and then do do max-pooling in each sub-window

Best way to draw an apple?

What is given?	What can be drawn?
memory +  blank paper	 apple, with a bite
 +  ground truth seen blank paper	 apple, with a bite on the right, a leaf on top
ground truth reference on paper 	 THE apple logo, **pixel-to-pixel aligned**

Drawbacks of RoI pooling for segmentation task

- RoIPool *breaks* pixel-to-pixel translation-equivariance

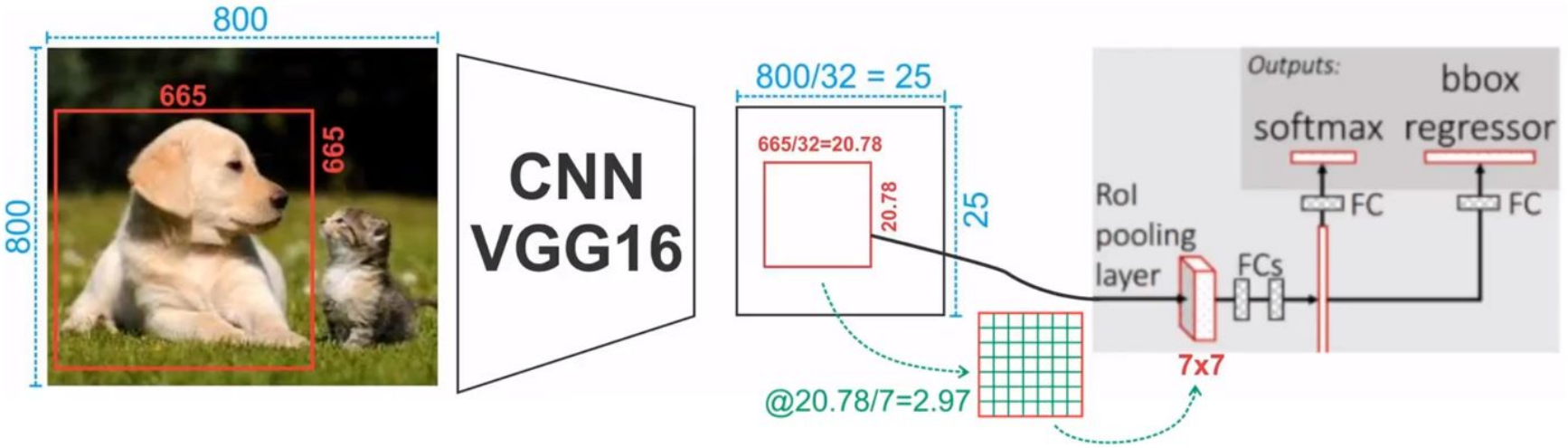


Translation equivariance: shifts in input lead to corresponding changes in the feature map

Source: author's slide

One pixel in RoI means many pixels in the raw image

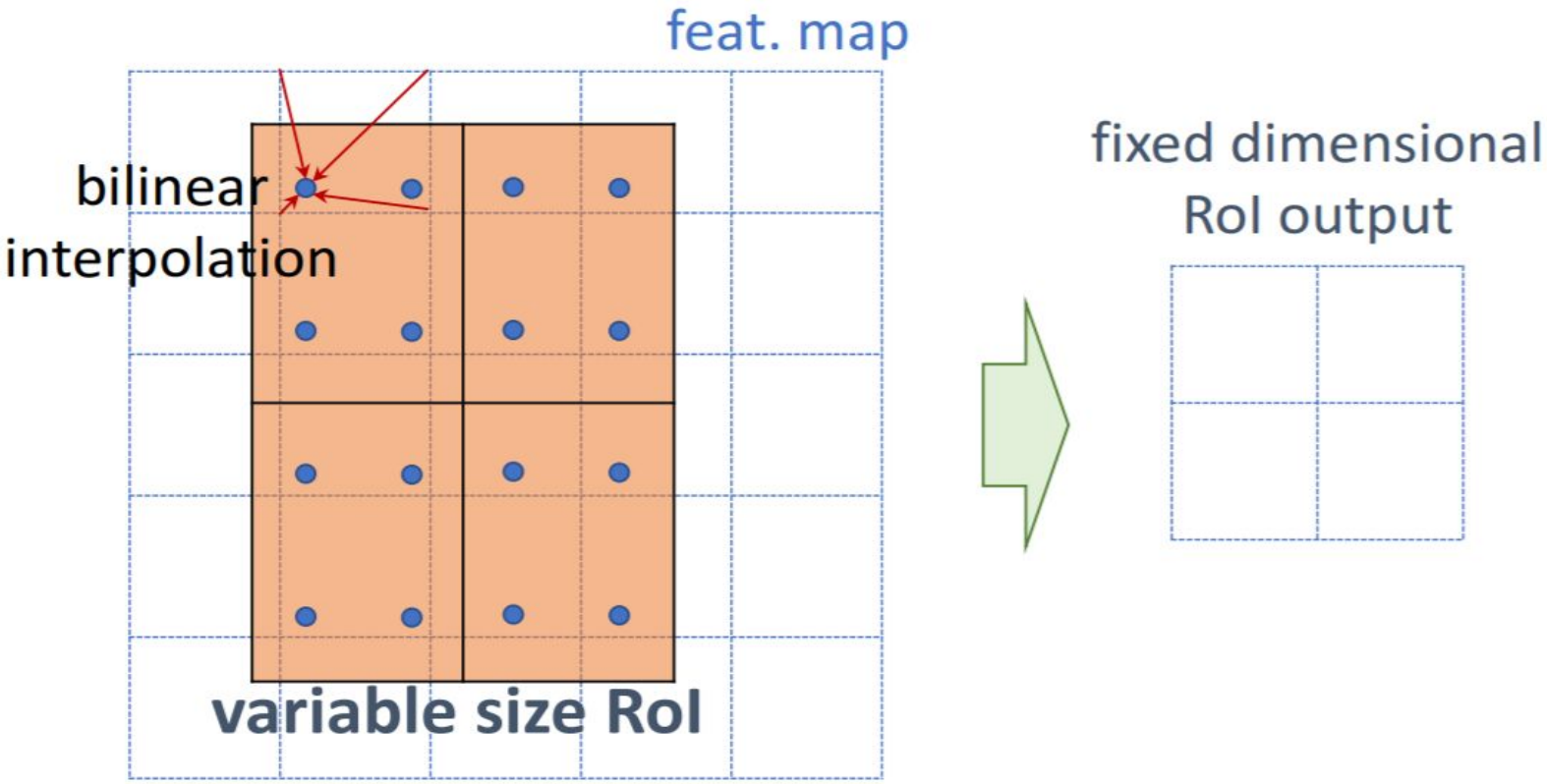
The **misalignment** introduced by **Quantization** associated with **RoIPool matters** for a **large stride**



Slide source: https://www.youtube.com/watch?v=XGi-Mz3do2s&t=288s

Solution: RoI align

Property 1: Pixel-to-pixel alignment



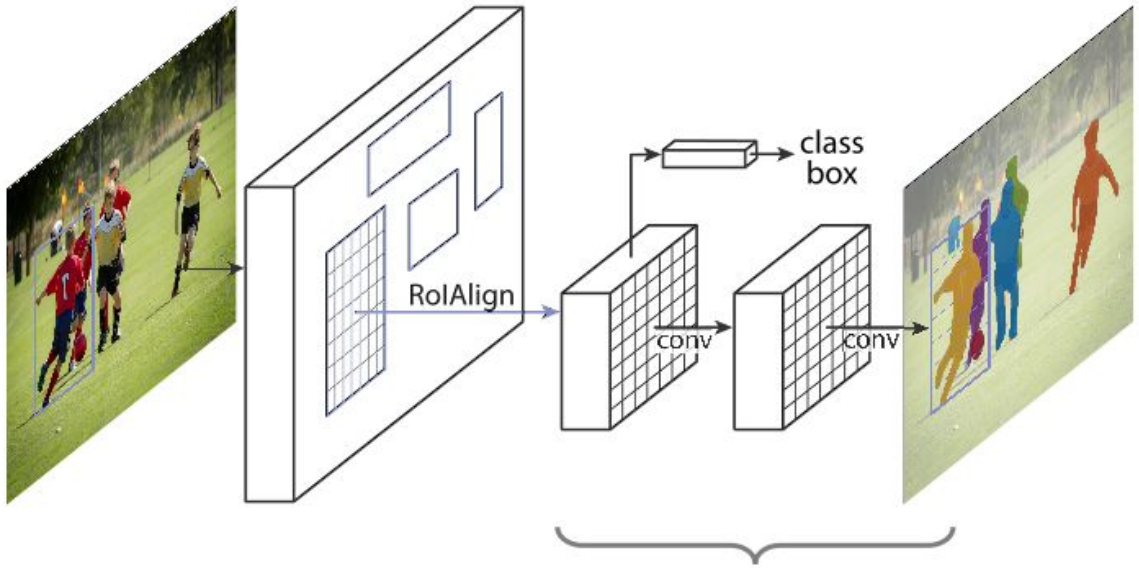
https://en.wikipedia.org/wiki/Bilinear_interpolation

Source: author's slide

Solution: RoI align

Property 2: equivariance

Equivariance in Mask R-CNN

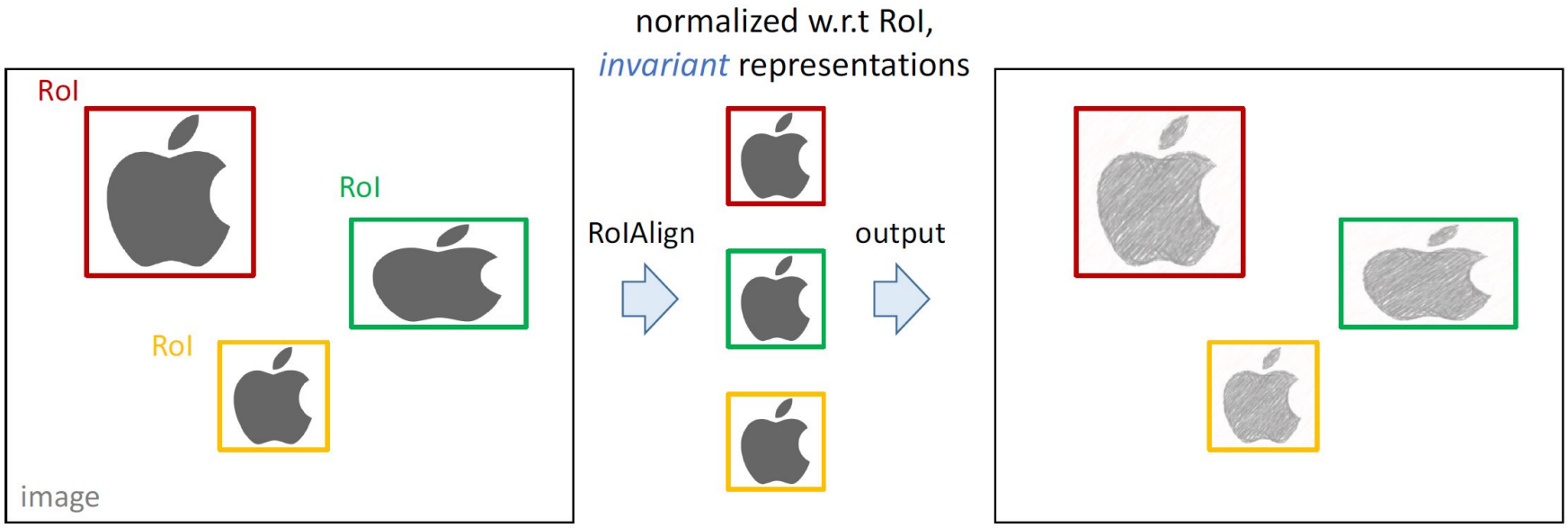


2. Fully-Conv on RoI:
equivariant to translation within RoI

Source: author's slide

Solution: RoI align

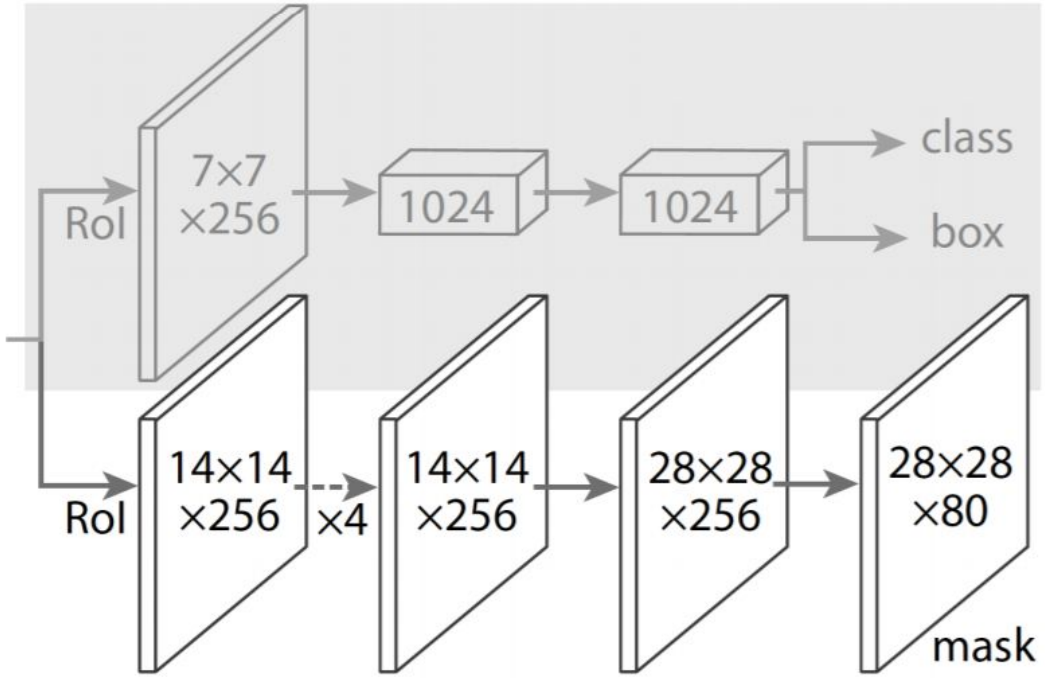
Property 2: Scale-invariant



- RoIAlign creates *scale-invariant* representations
- RoIAlign + "output pasted back" provides *scale-equivariance*

Source: author's slide

FCN Mask head



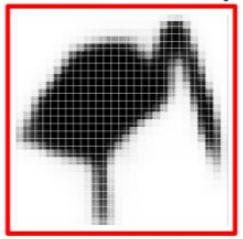
Per class **binary masks** are generated rather than a **multinomial mask**. No class competition.

How is a mask generated from FCN head?

- Pixel-to-pixel aligned

RoI

28x28 FCN prediction

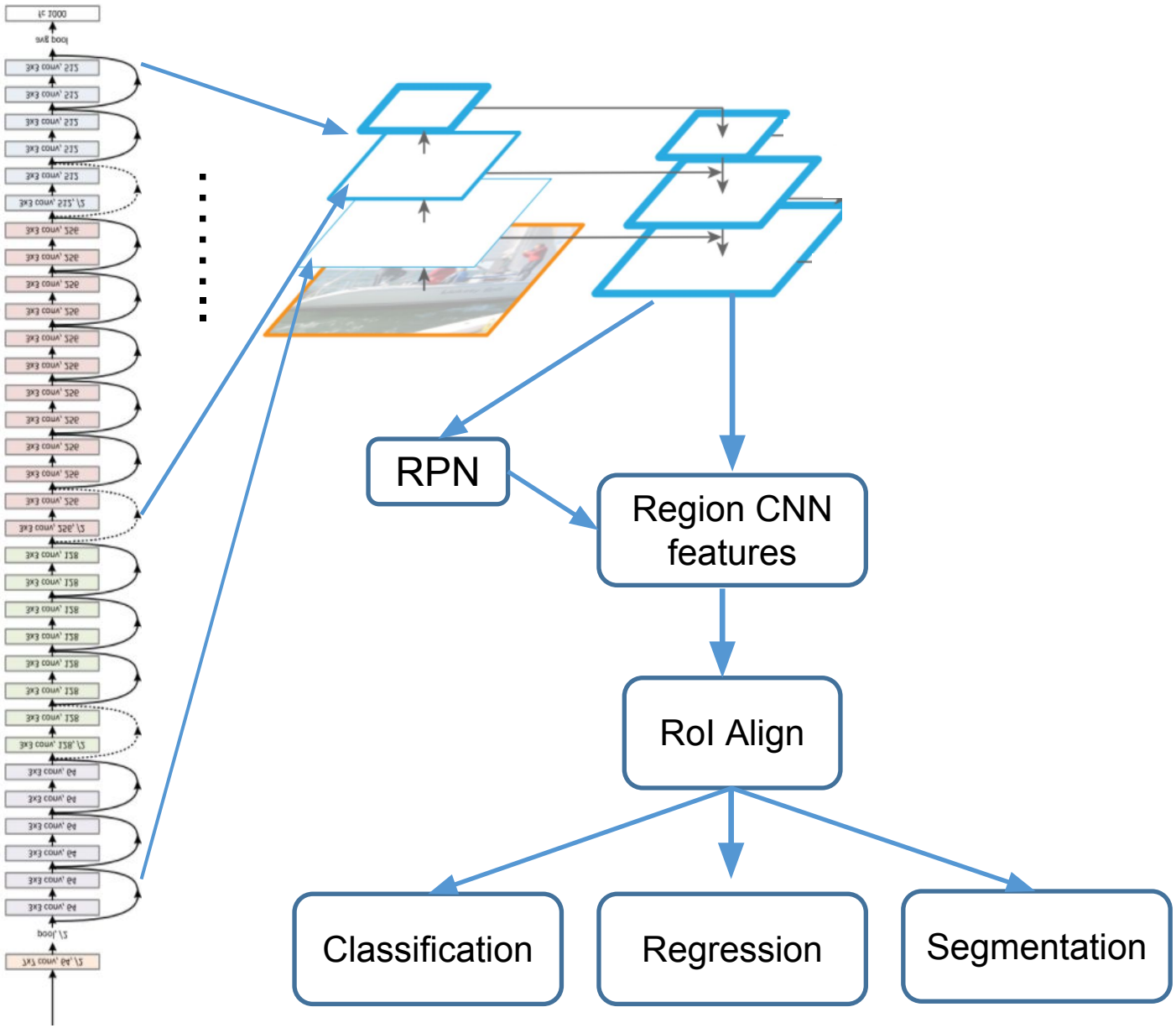


resized soft prediction

final mask

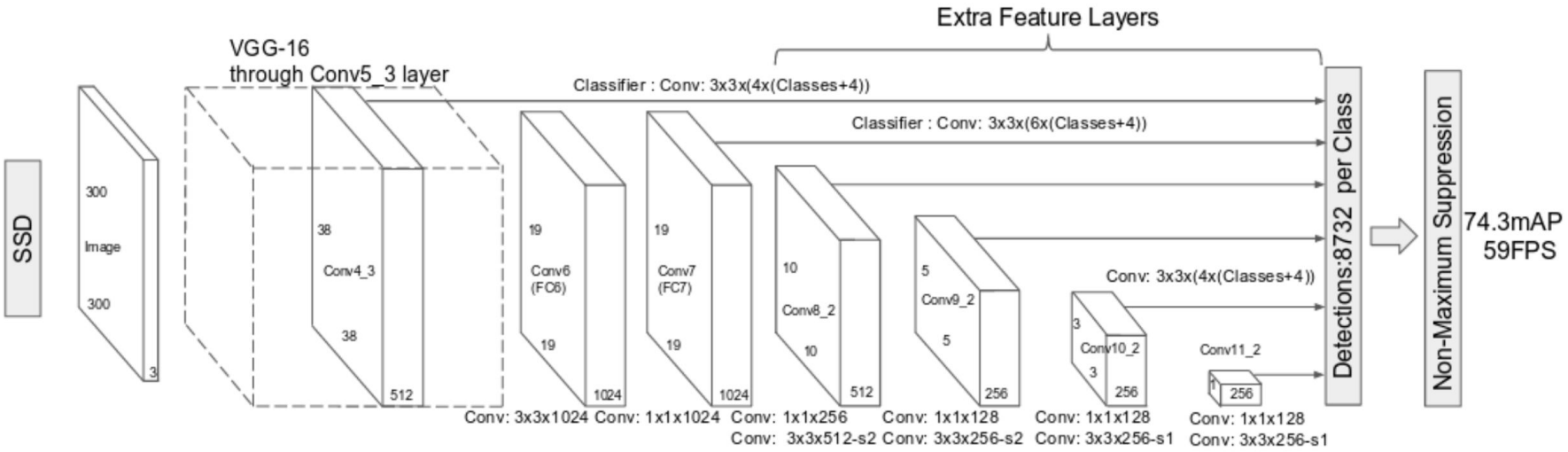
Source: author's slide

Backbone of Mask-rcnn: FPN

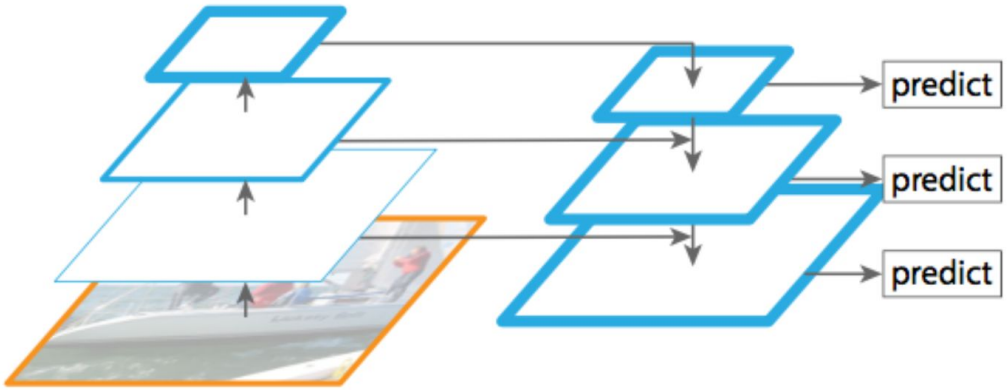


Backbone of Mask-rcnn: FPN

SSD: skip connection



FPN: lateral connection

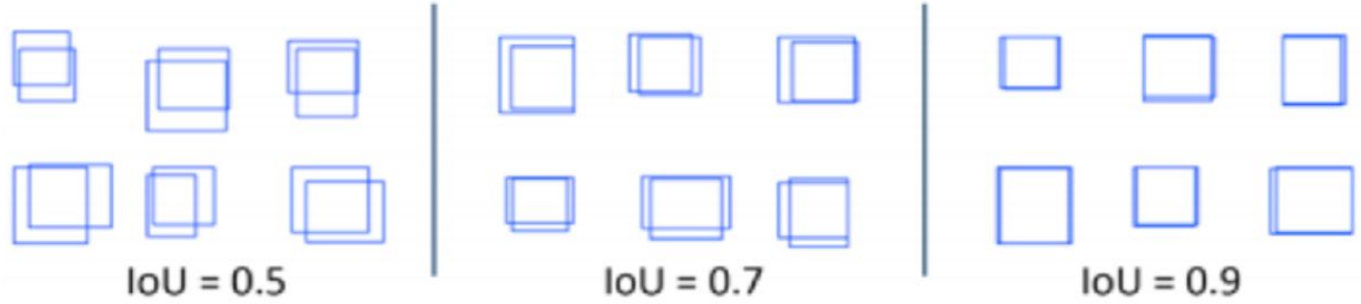


Experiments

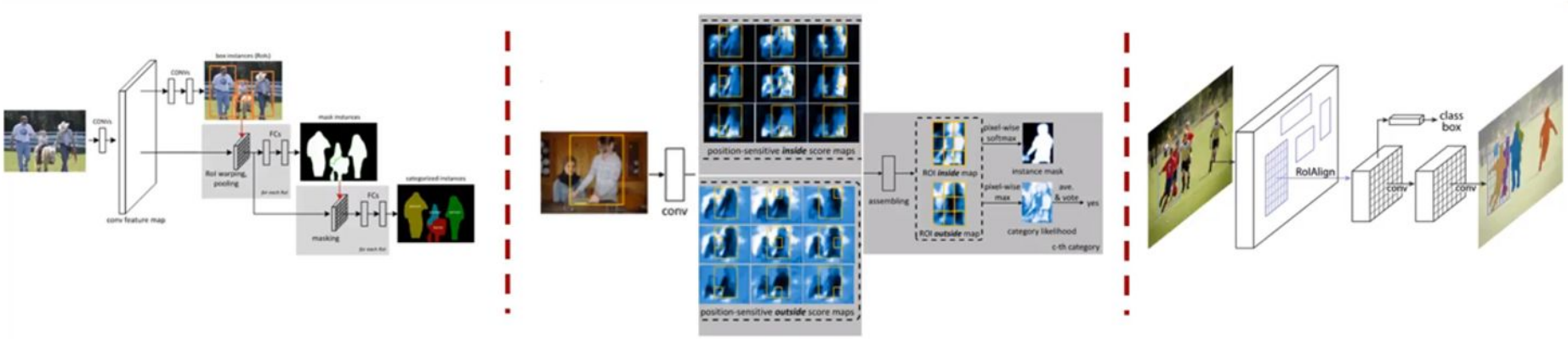
- Main dataset: MS COCO
  - 80 classes
  - 80k train image
  - 35k sub– set of val images
  - 5k images for ablation experiments
- Metric

Challenges Score: AP

- AP is averaged over multiple IoU values between 0.5 and 0.95.
- More comprehensive metric than the traditional AP at a fixed IoU value (0.5 for PASCAL).



State-of-art in 2016: FCIS



	MNC	FCIS	Mask R-CNN
Mask	All classes are in one mask (translation-invar.)	Position sensitive mask (translation-var.) [1]	Binary mask for each object
Feature map	RoI-Warping	?	RoIAlign
Input classification	Depend mask	Join mask pred. & classi.	Parallel-independent

Main results: Mask R-CNN vs MNC and FCIS

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

Ablation: different 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**

(a) **Backbone Architecture**: Better backbones bring expected gains: deeper networks do better, FPN outperforms C4 features, and ResNeXt improves on ResNet.

More efficient encoder and decoder help generally.

Ablation: Binary vs Multinomial Loss

	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(b) **Multinomial *vs*. Independent Masks** (ResNet-50-C4): *Decoupling* via per-class binary masks (sigmoid) gives large gains over multinomial masks (softmax).

Ablation: RoIAlign vs RoPool

baseline: ResNet-50-Conv5 backbone, **stride=32**

	mask AP			box AP		
	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*

- huge gain at high IoU, in case of big stride (32)

Bounding box detection results are improved by RoIAlign

ResNeXt-101-FPN on COCO.

	backbone	AP^{bb}	AP^{bb}_{50}
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6

bbox improved by:

- RoIAlign

Bounding box detection benefits from mult-task learning

Bounding box detection benefits from the segmentation task

Backbone: ResNeXt-101-FPN on COCO.

	backbone	AP^{bb}	AP^{bb}_{50}
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6
Mask R-CNN	ResNet-101-FPN	38.2	60.3

bbox improved by:

- RoIAlign
- Multi-task training w/ mask

Qualitative Results on Instance Segmentation

https://www.youtube.com/watch?v=g7z4mkfRjI4&t=639s

Generality of Mask-rcnn: Human keypoint detection

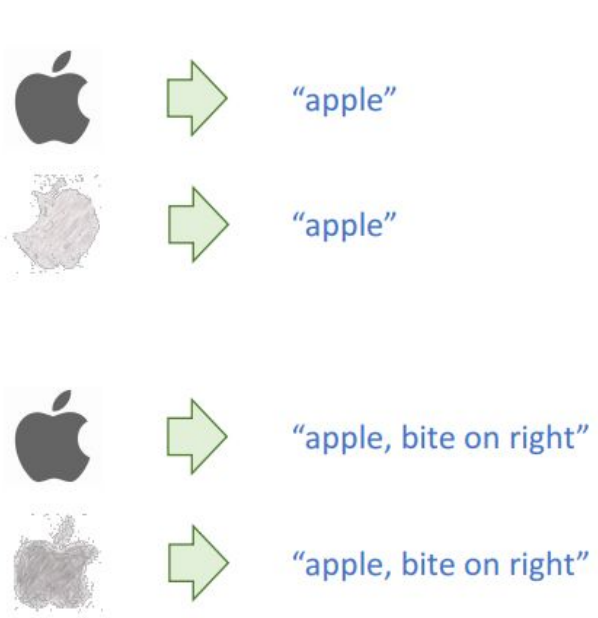
- keypoint = 1-hot mask
- human pose = 17 masks

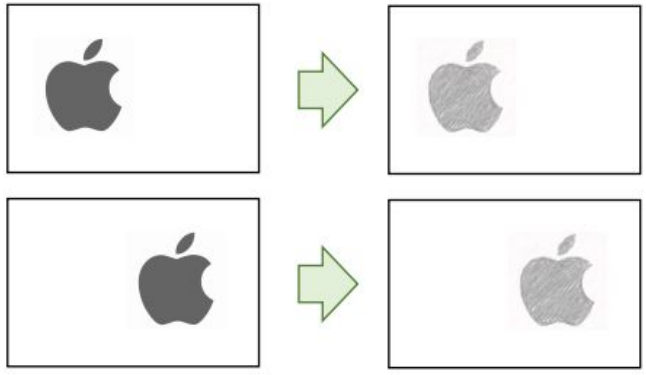
	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
CMU-Pose+++ [6]	61.8	84.9	67.5	57.1	68.2
G-RMI [32]†	62.4	84.0	68.5	**59.1**	68.1
Mask R-CNN, keypoint-only	62.7	87.0	68.4	57.4	71.1
Mask R-CNN, keypoint & mask	**63.1**	**87.3**	**68.7**	57.8	**71.4**

Thank you!

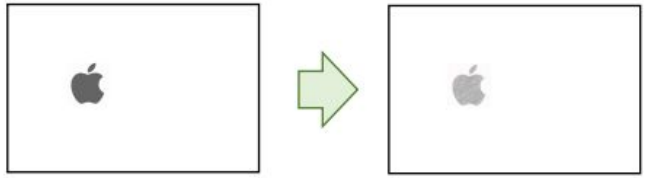
Supplementary slides follow

Invariance vs. Equivariance





translation-equivariant



scale-equivariant

see also "What is wrong with convolutional neural nets?", Geoffrey Hinton, 2017

Invariance vs equivariance

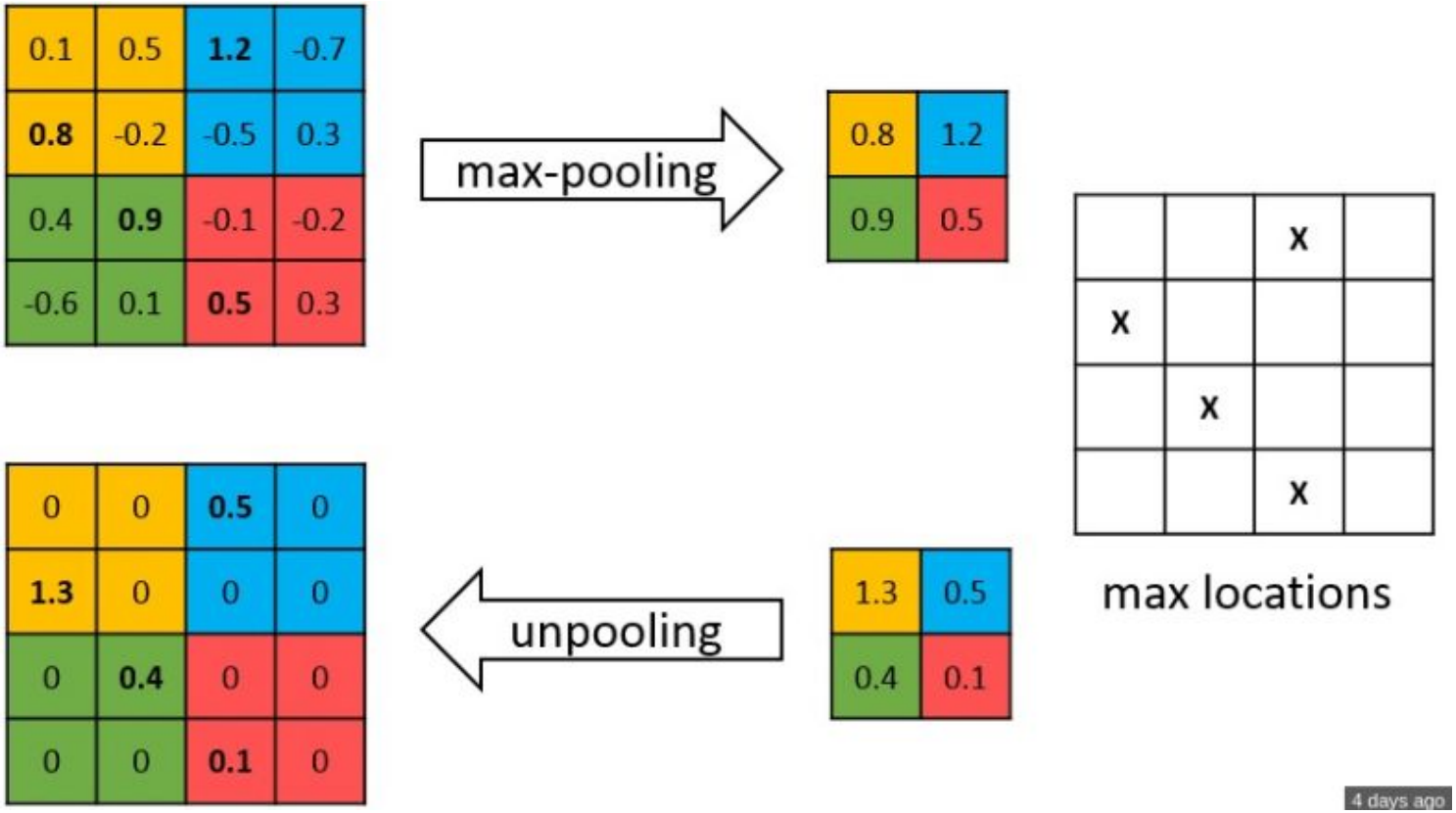
Translation means that each point/pixel in the image has been moved the same amount in the same direction

Convolution operation is translation equivariant
Maxpooling is translation invariance

Ref:

Method for upsampling in segmentation task

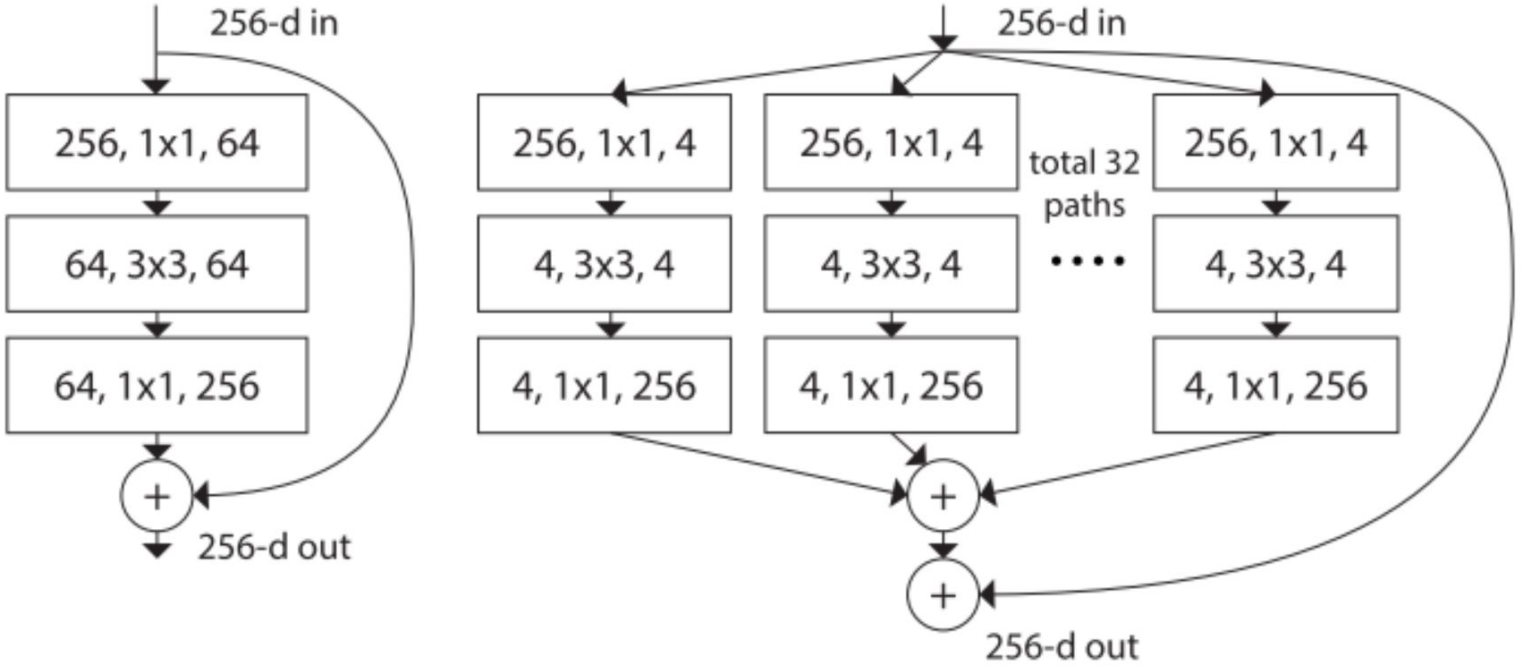
1. **Uppooling**
2. **Interpolation**
3. **deconvolution**



Ref:

ResNeXt: what is new compared to Resnet?

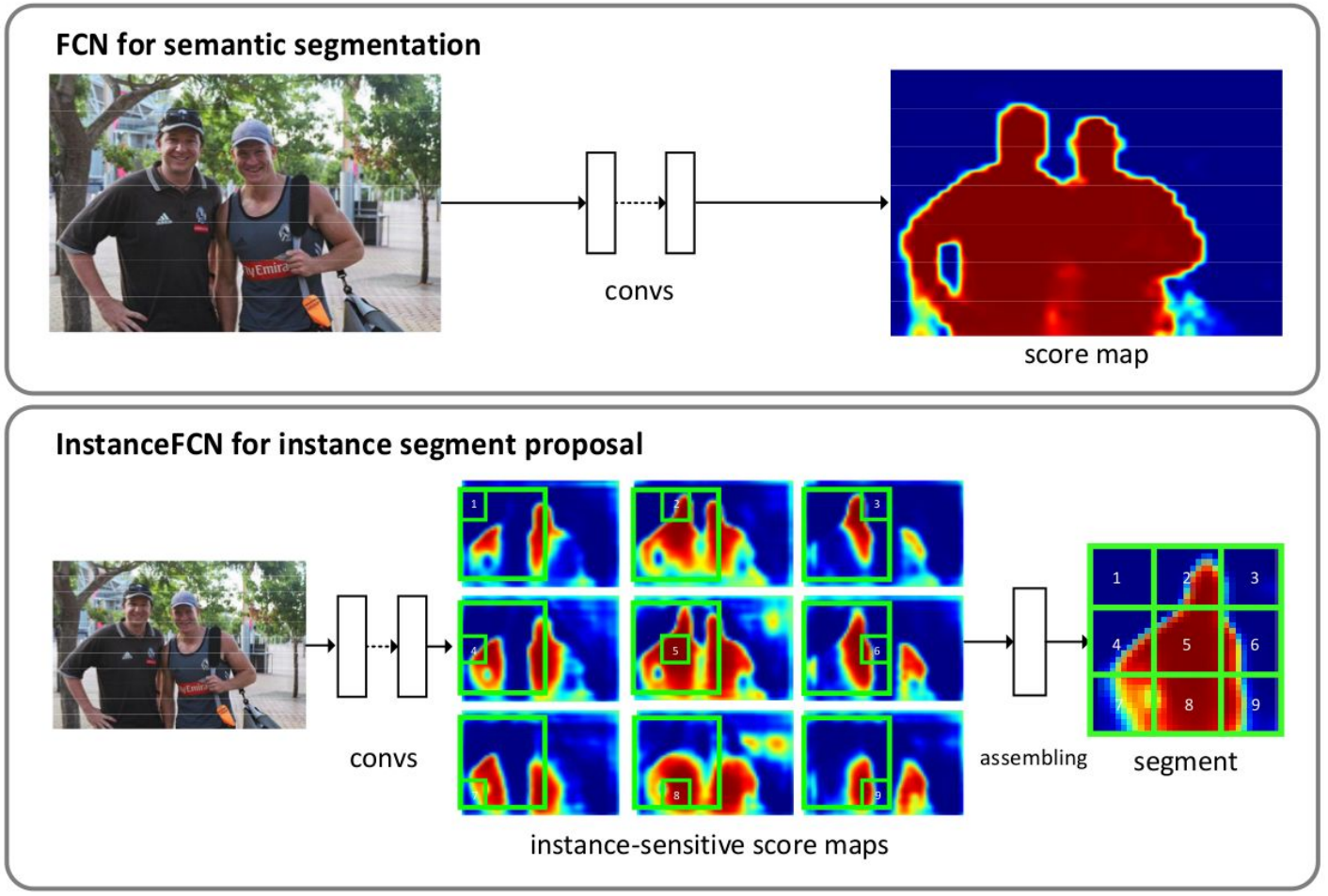
Similar to Inception net but **share the same topology** among the multiple paths and use **summation rather than concatenation** to merge



Cardinality—the number of independent paths
Width: the number of channels in output.
Note: convolution is 3d computation.

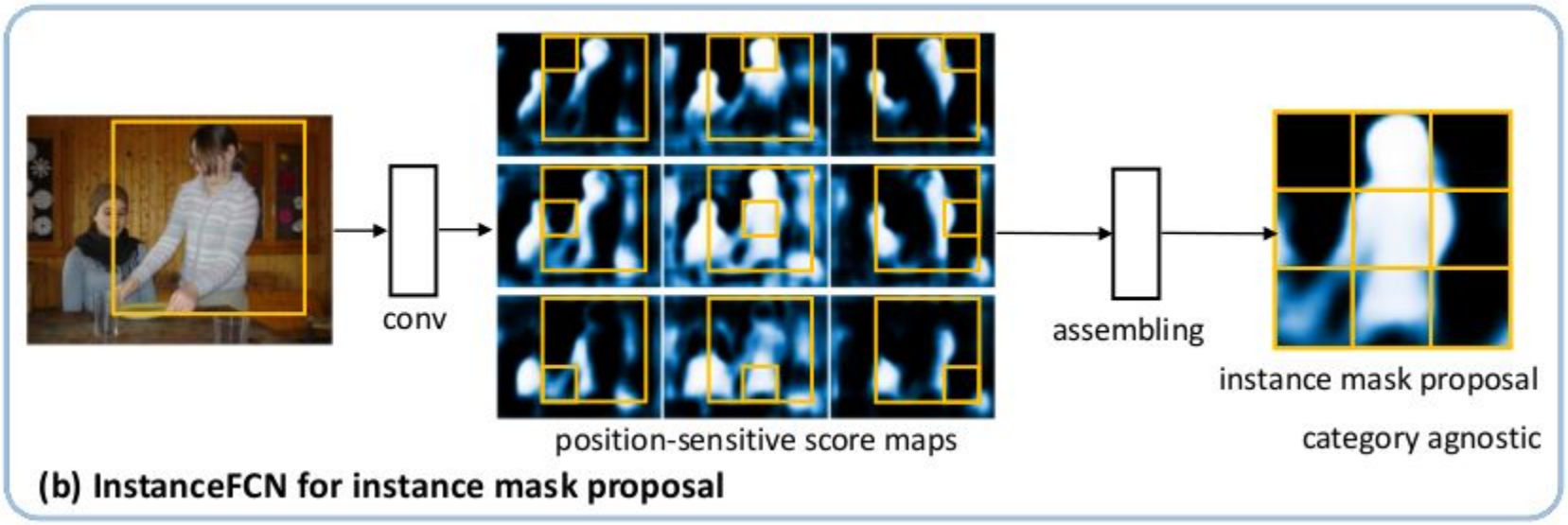
Ref: S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He. Aggregated Residual Transformations for Deep Neural Networks

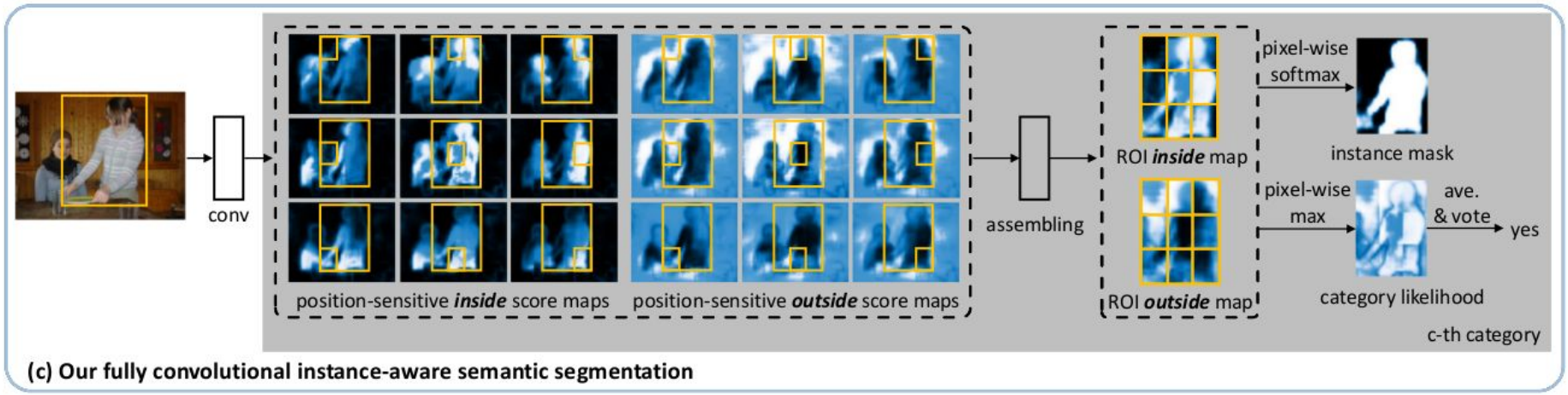
Position sensitive score map



Ref: Jifeng Dai, et al. Instance FCN

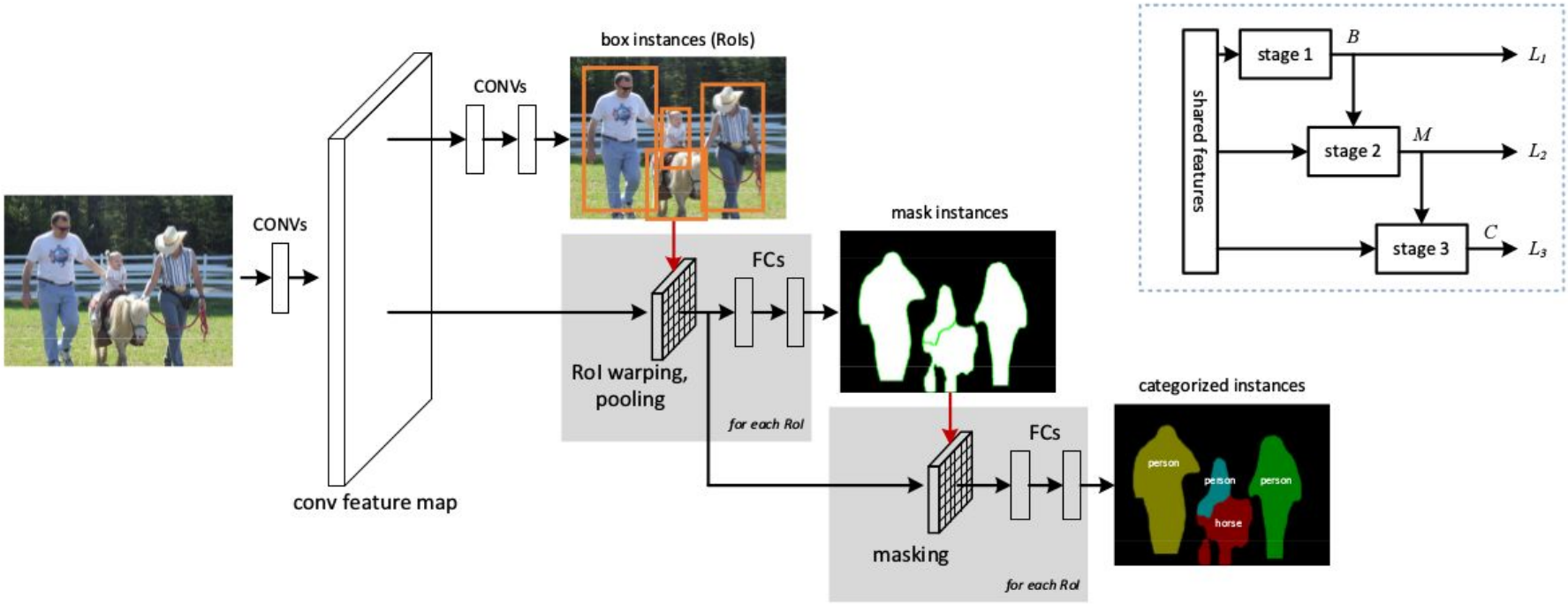
Instance FCN vs FCIS





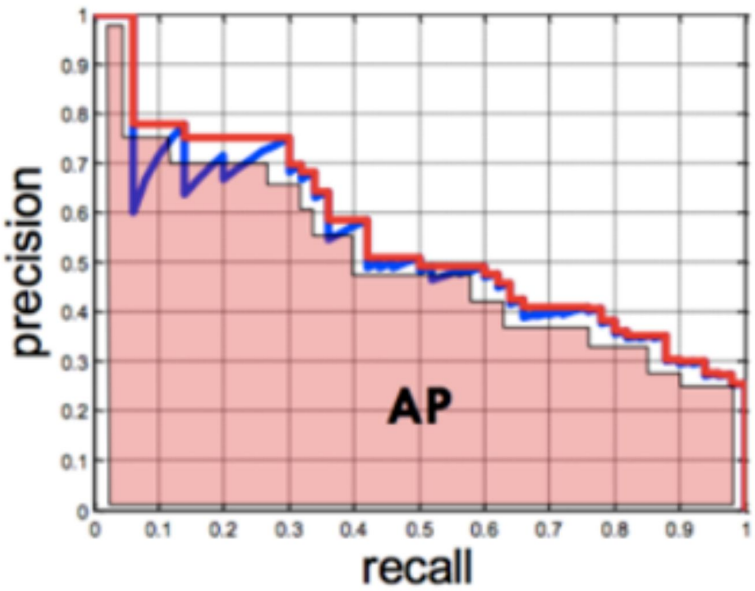
Ref: Yi Li, et al. FCIS

Multi-task Network Cascades



Ref: Yi Li, et al. Instance-aware Semantic Segmentation via Mult-task Network Cascades

AP metric



Commonly, IoU > 0.5 means that it was a hit, otherwise it was a fail. For each class, one can calculate the

- True Positive ($TP(c)$): a proposal was made for class c and there actually was an object of class c
- False Positive ($FP(c)$): a proposal was made for class c, but there is no object of class c
- Average Precision for class c: $\frac{\#TP(c)}{\#TP(c)+\#FP(c)}$

The mAP (mean average precision) = $\frac{1}{|classes|}\sum_{c\in classes}\frac{\#TP(c)}{\#TP(c)+\#FP(c)}$

Fig source: http://darkpgmr.tistory.com/162

RoI warp vs RoIPool

The RoI warping layer crops a feature map region and warps it into a target size by interpolation.

Ref: Jifeng Dai, kaiming He, Jian Sun Instance-aware Semantic Segmentation via Multi-task Network Cascades

Timing

ResNet-50-FPN on COCO trainval135k

8-GPU

0.72s per 16-image mini-batch