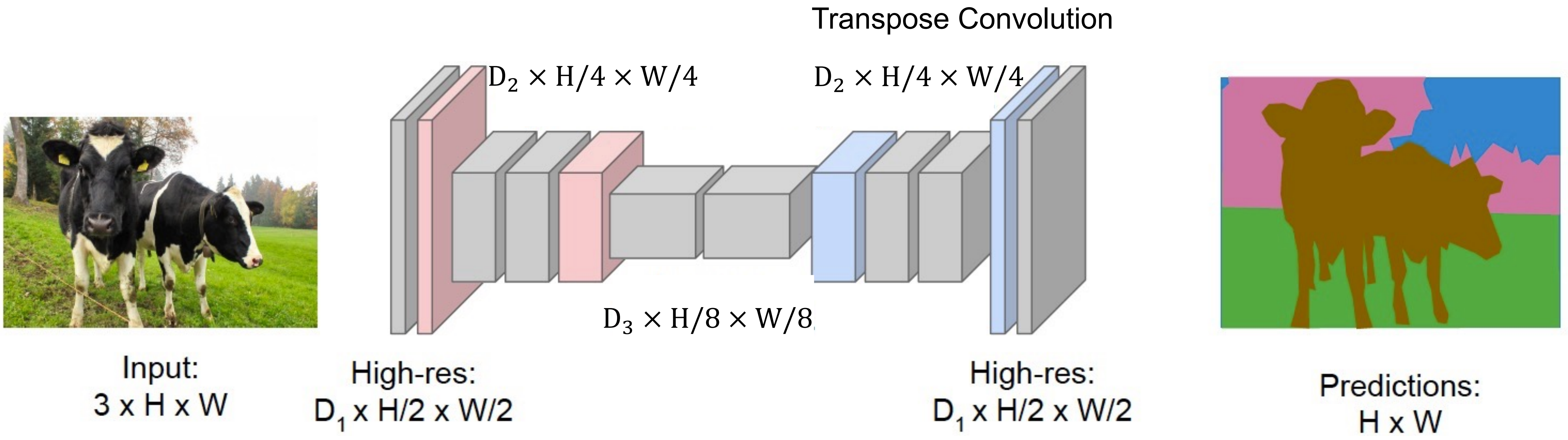
Object Detection 2

Xiaolong Wang

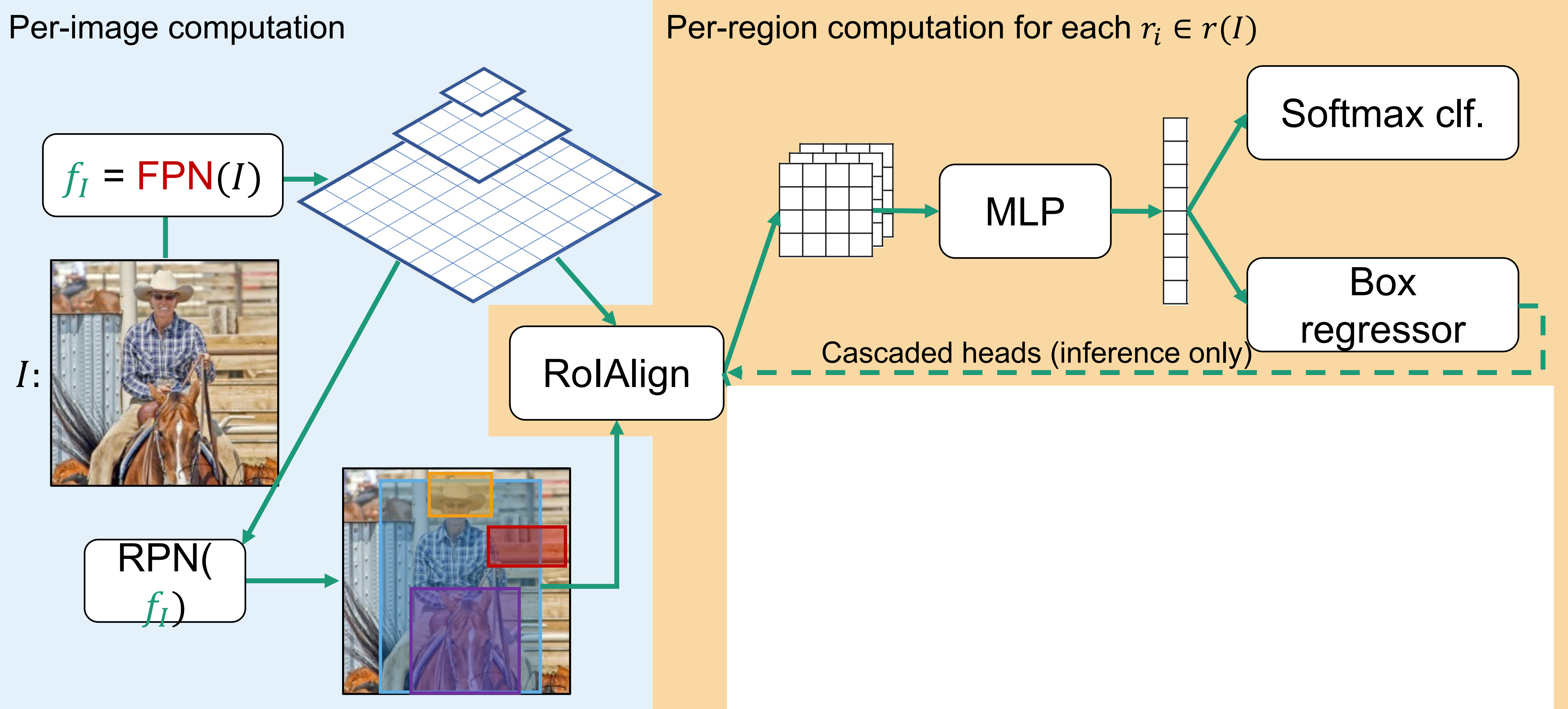
Mask R-CNN

Fully Convolutional Network (FCN)

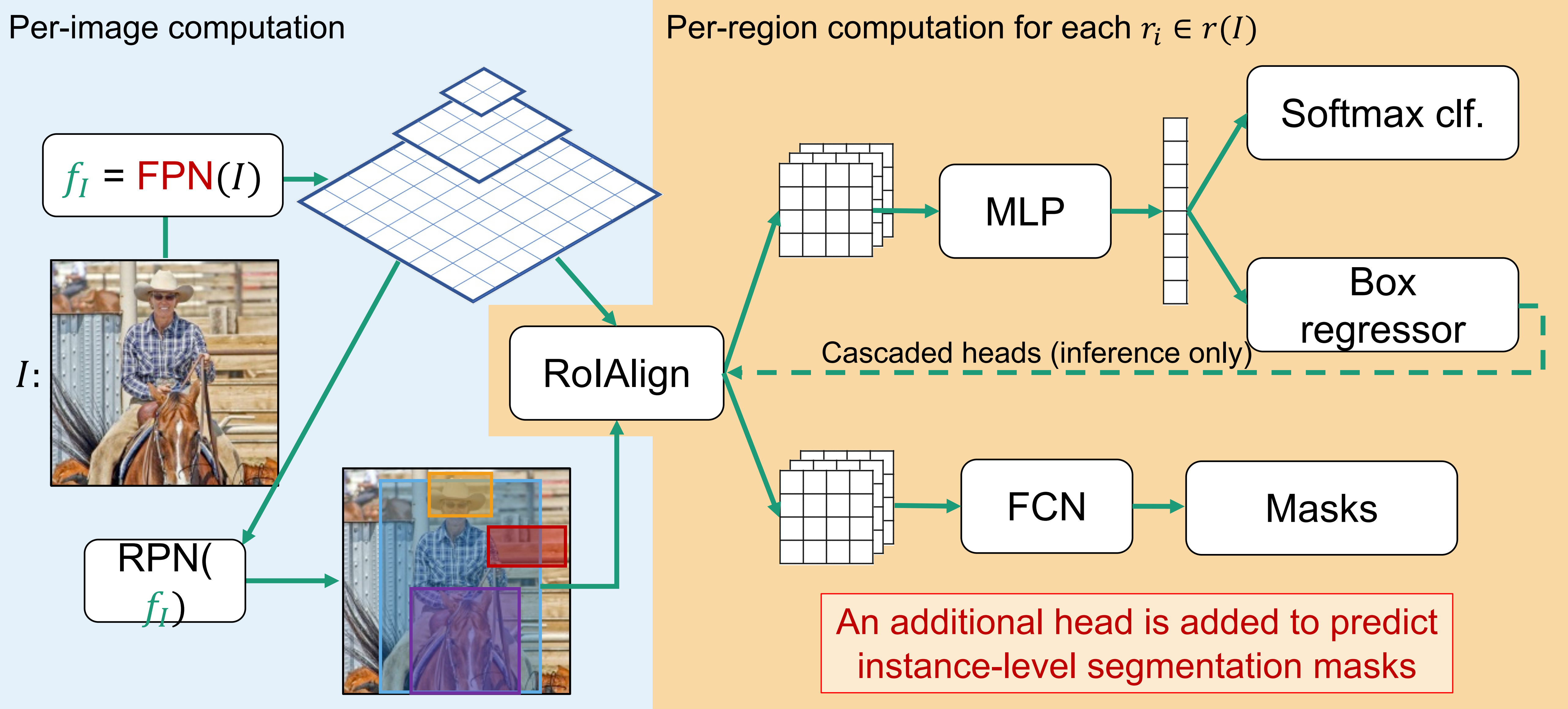


He et al. Mask R-CNN. ICCV 2017.

Mask R-CNN

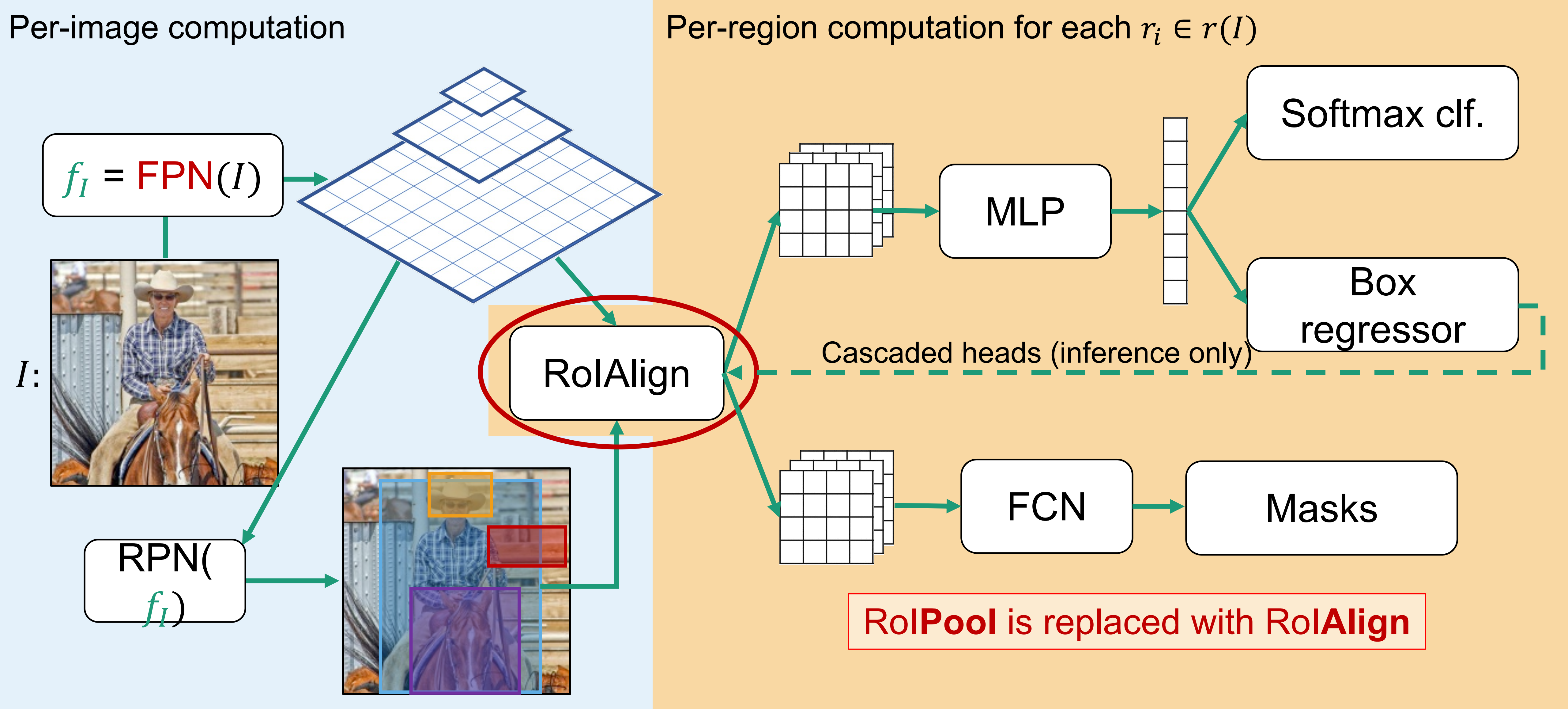


Mask R-CNN



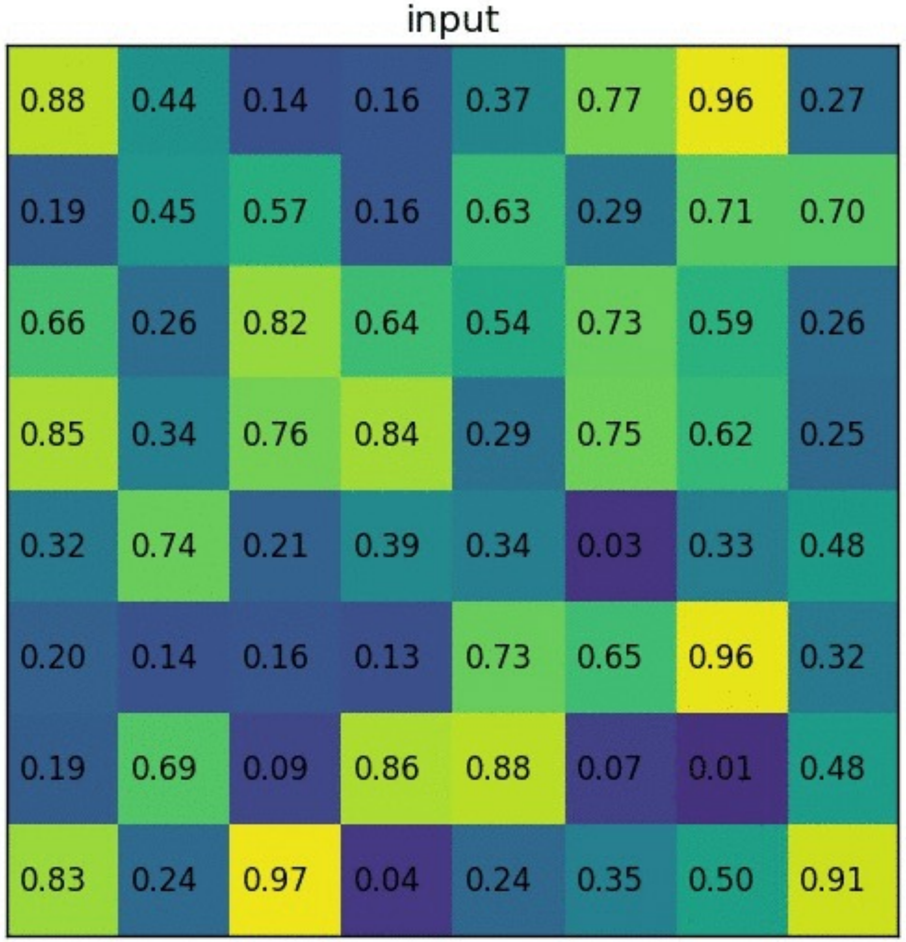
An additional head is added to predict instance-level segmentation masks

Mask R-CNN



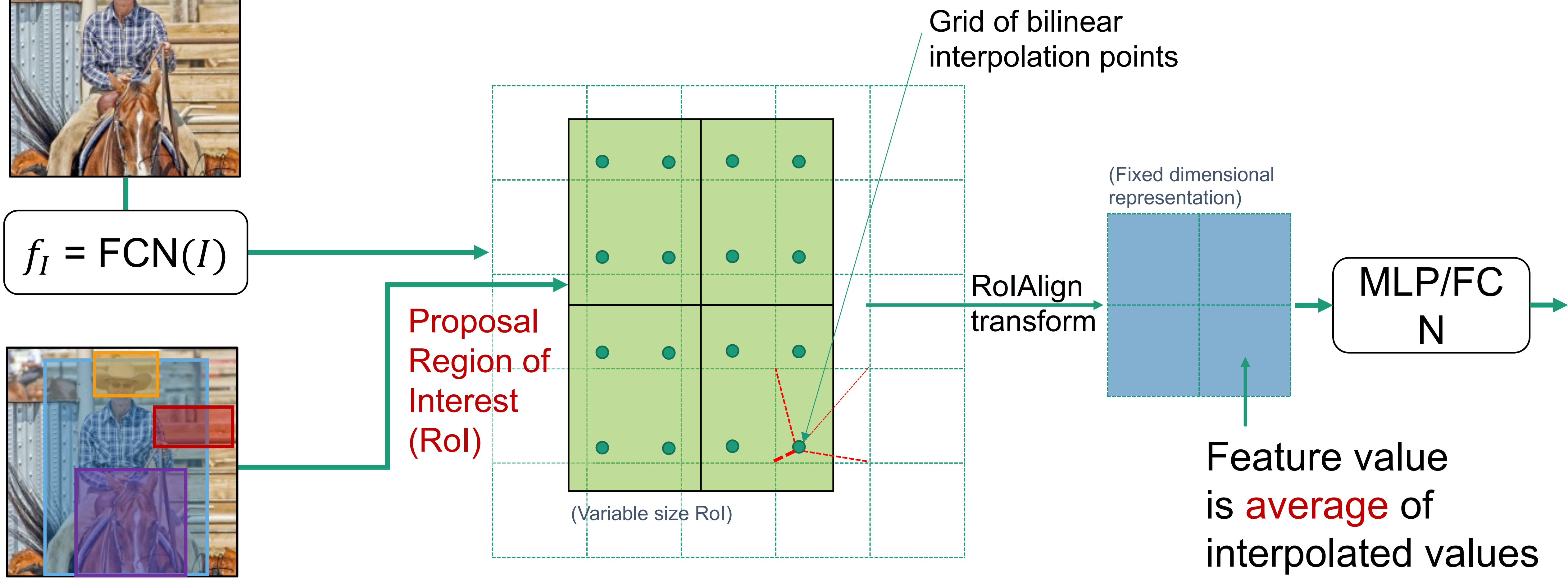
Per-image computation
Per-region computation for each $r_i \in r(I)$
f_I = FPN(I)
I:
RPN(f_I)
RoIAlign
MLP
Softmax clf.
Box regressor
Cascaded heads (inference only)
FCN
Masks
RoI**Pool** is replaced with RoI**Align**

RoI pooling

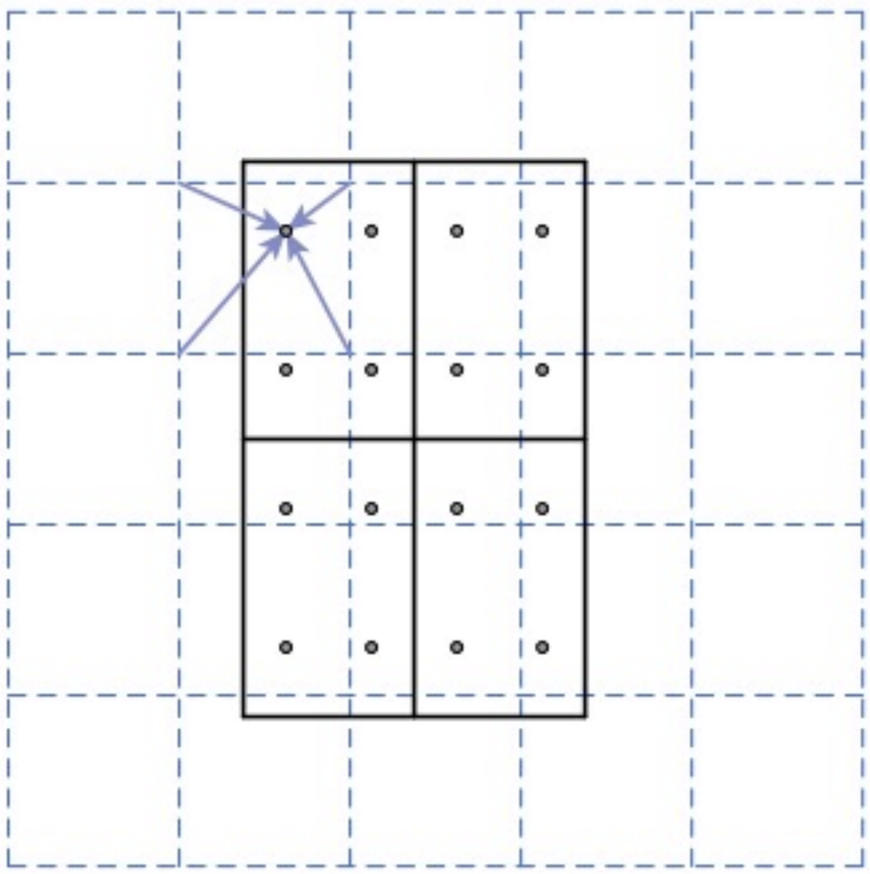

input
0.88 0.44 0.14 0.16 0.37 0.77 0.96 0.27
0.19 0.45 0.57 0.16 0.63 0.29 0.71 0.70
0.66 0.26 0.82 0.64 0.54 0.73 0.59 0.26
0.85 0.34 0.76 0.84 0.29 0.75 0.62 0.25
0.32 0.74 0.21 0.39 0.34 0.03 0.33 0.48
0.20 0.14 0.16 0.13 0.73 0.65 0.96 0.32
0.19 0.69 0.09 0.86 0.88 0.07 0.01 0.48
0.83 0.24 0.97 0.04 0.24 0.35 0.50 0.91

RoI pooling → RoIAlign

Transform arbitrary size proposal into a fixed-dimensional representation (e.g., 2x2)

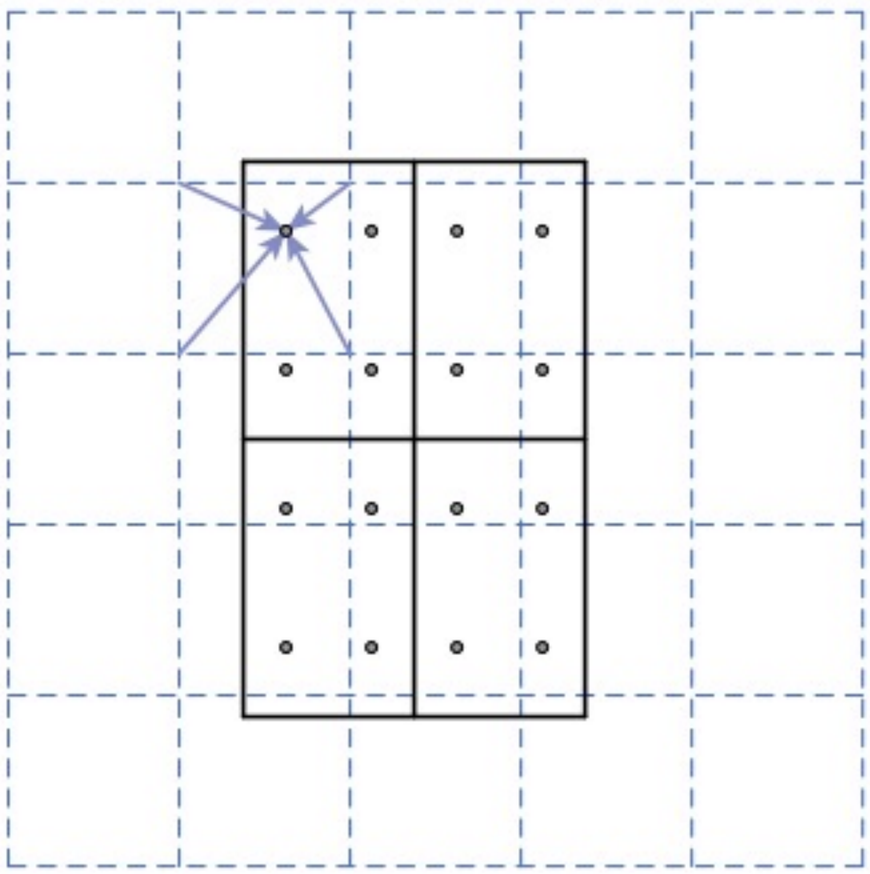


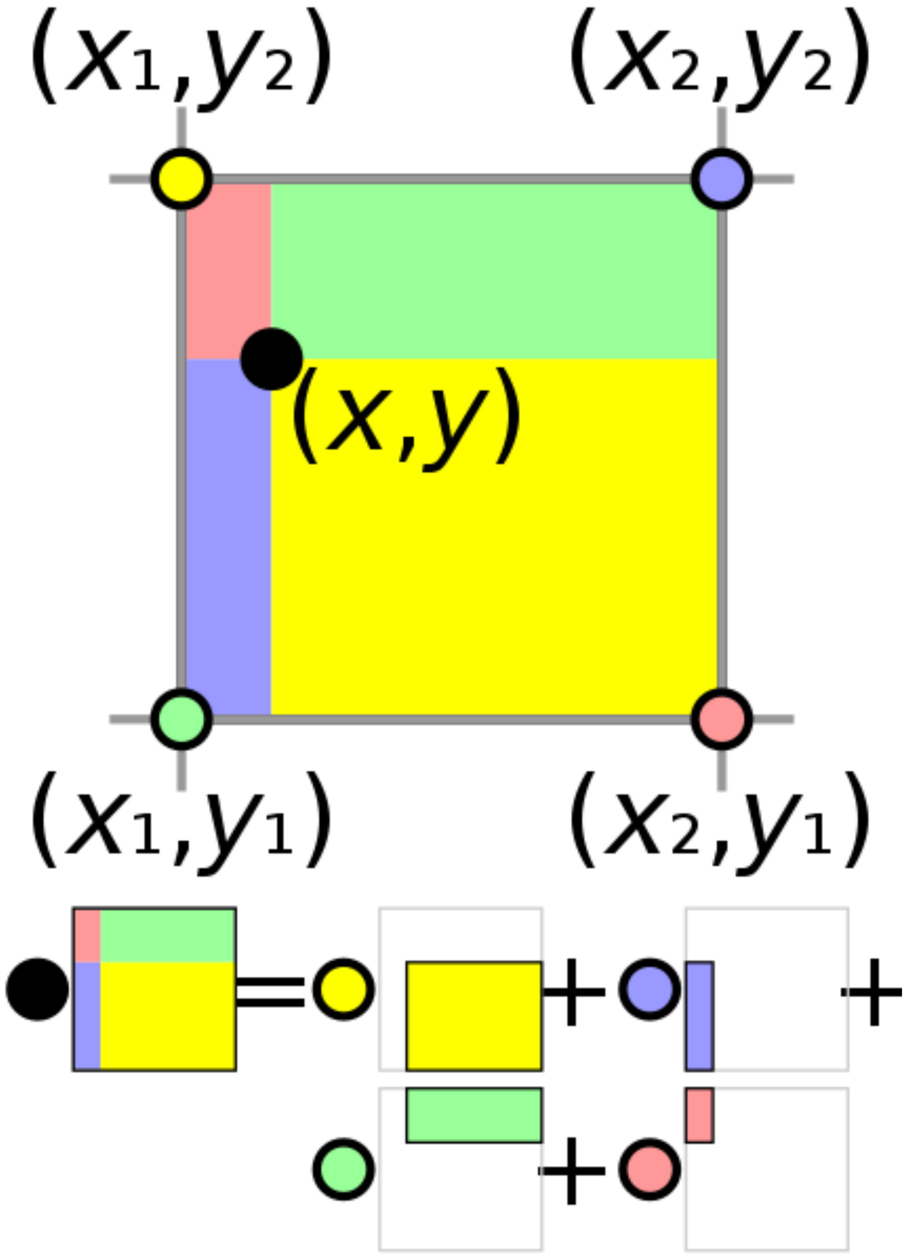
RoIAlign



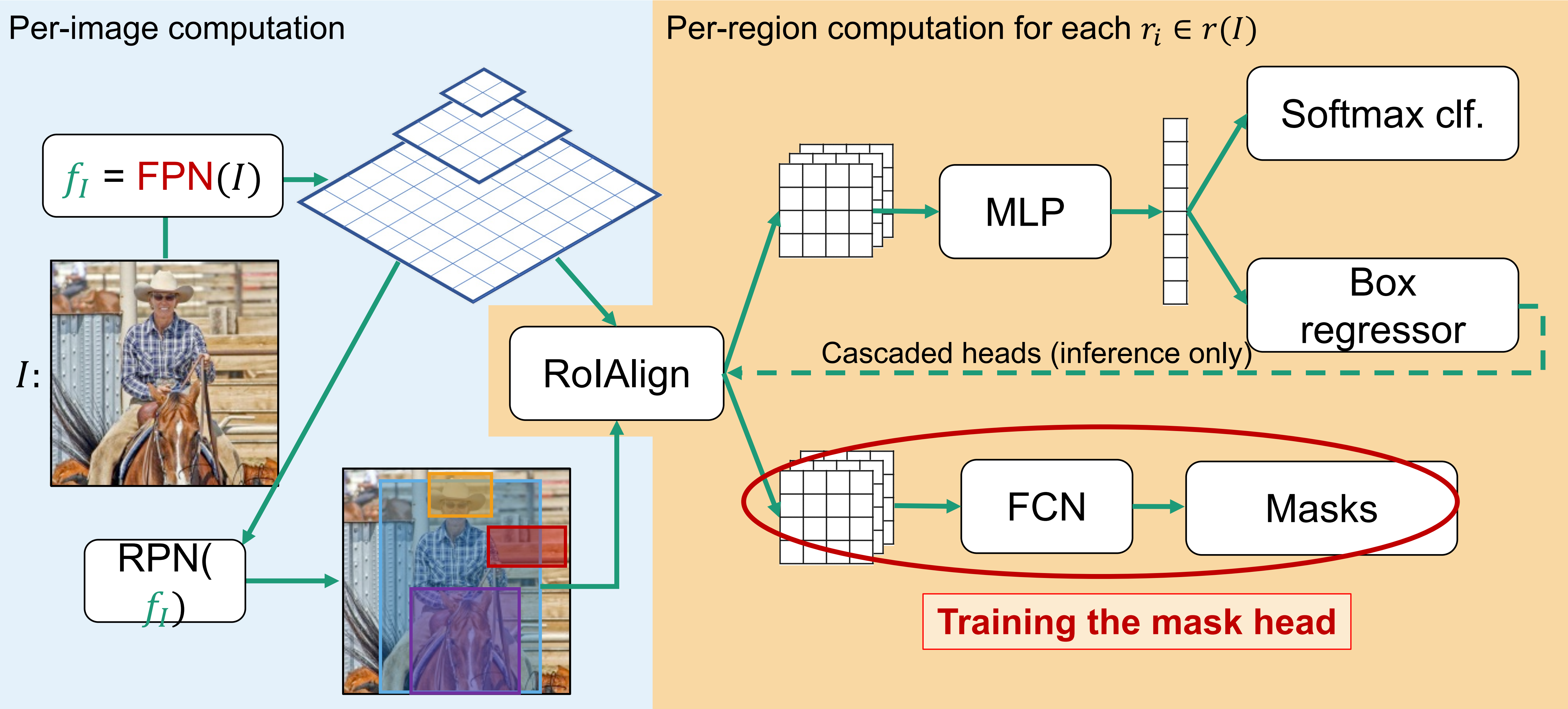
- Bilinear interpolation for each sampled location
- Use max pooling / avg pooling for each roi bin

RoIAlign





Mask R-CNN

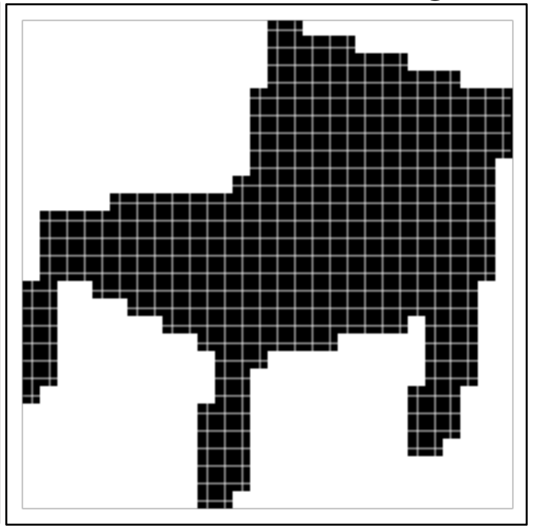


Per-image computation
Per-region computation for each $r_i \in r(I)$
f_I = FPN(I)
I:
RPN(f_I)
RoIAlign
MLP
Softmax clf.
Box regressor
Cascaded heads (inference only)
FCN
Masks
Training the mask head

Example Mask Training Targets

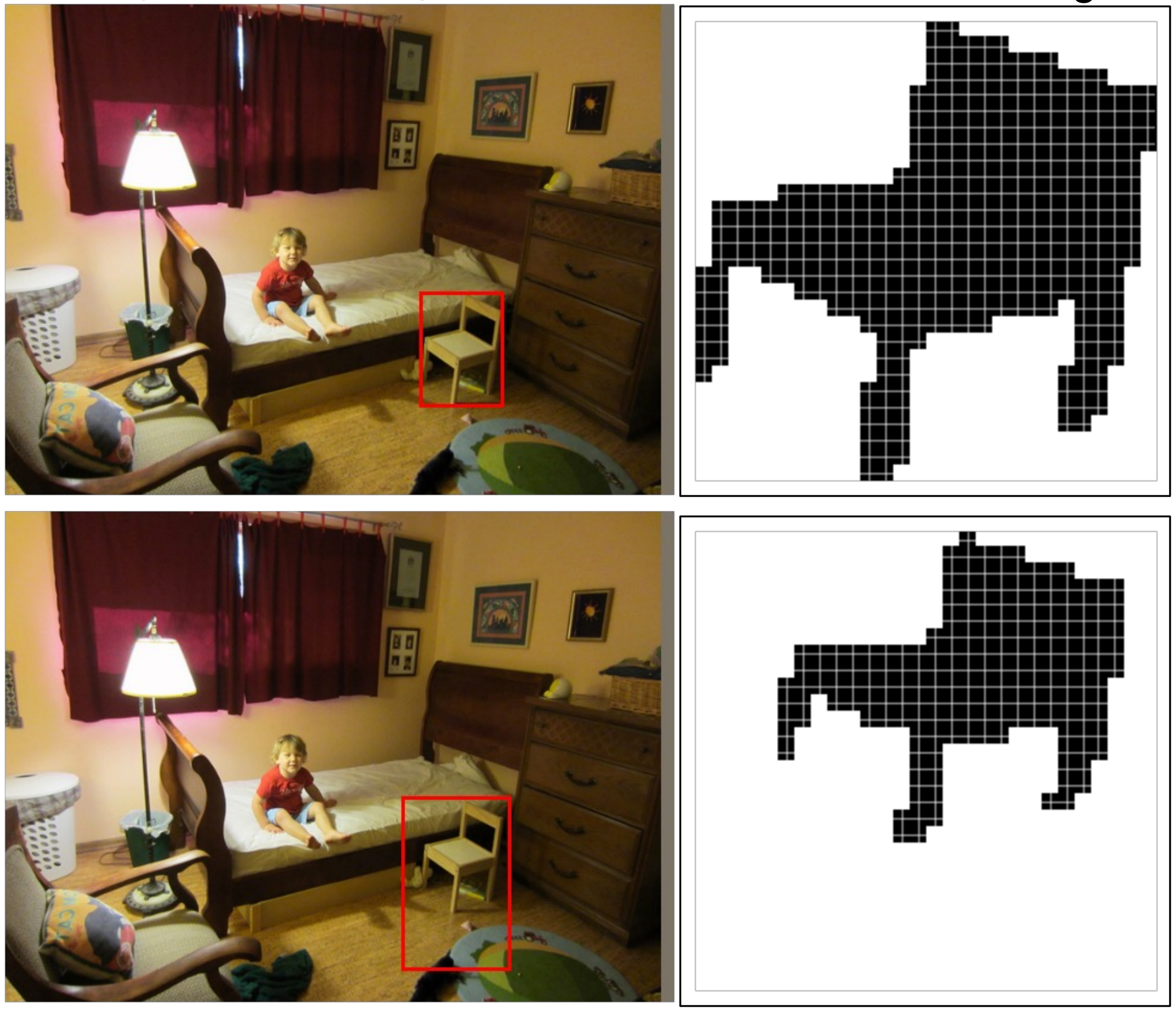
Image with training proposal

28x28 mask target



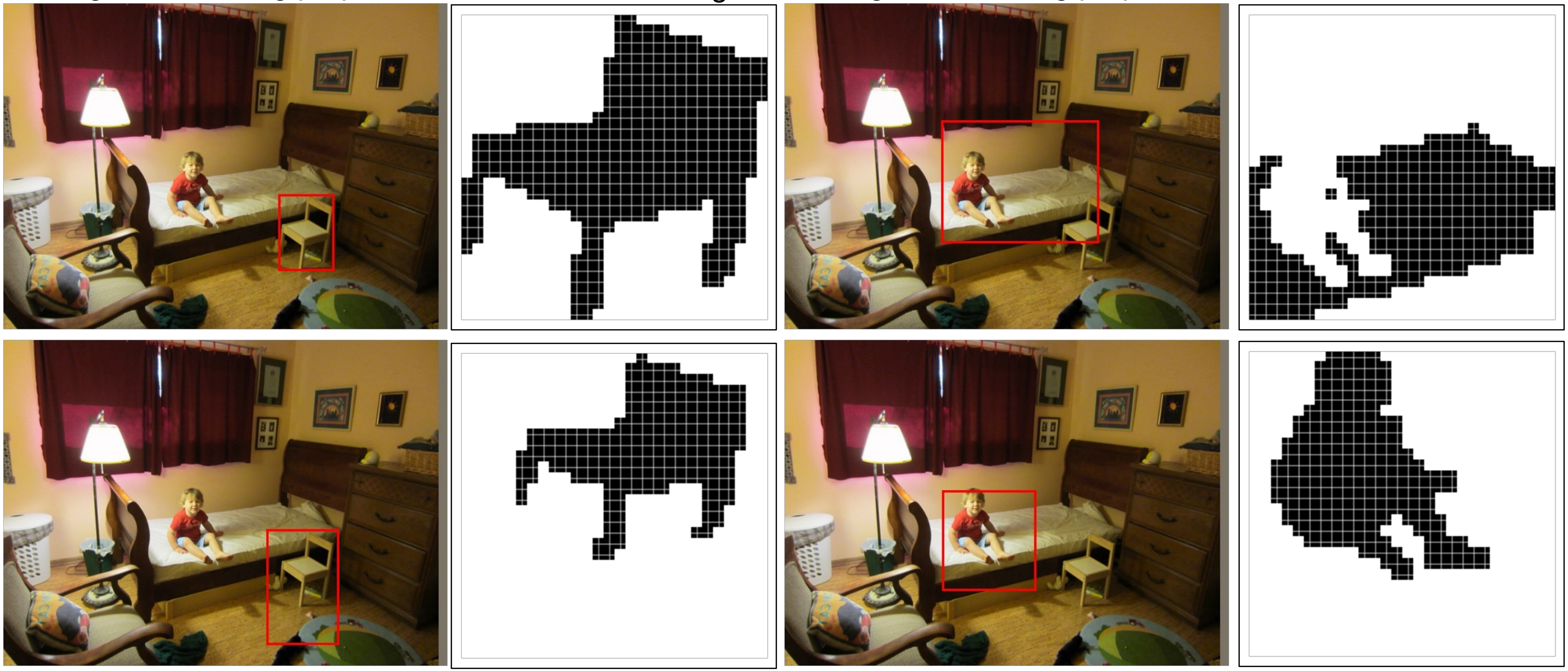
Example Mask Training Targets

Image with training proposal 28x28 mask target



Example Mask Training Targets

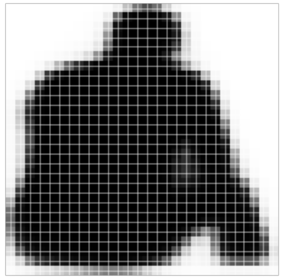
Image with training proposal | 28x28 mask target | Image with training proposal | 28x28 mask target



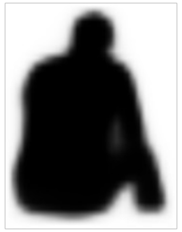
Binary Cross Entropy Loss on each pixel

Validation image with box detection shown in red

28x28 soft prediction from Mask R-CNN
(enlarged)



Soft prediction resampled to image coordinates
(bilinear and bicubic interpolation work equally well)

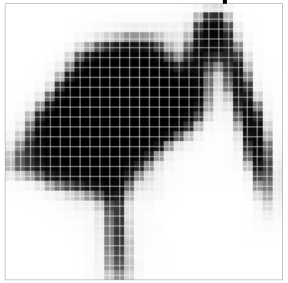


Final prediction (threshold at 0.5)

Binary Cross Entropy Loss on each pixel

Validation image with box detection shown in red

28x28 soft prediction



Resized Soft prediction

Final mask

person1.00
person1.00
person1.00
person1.00
person.98
surfboard1.00
person.91
surfboard1.00
surfboard.98
surfboard1.00
person.74
surfboard1.00

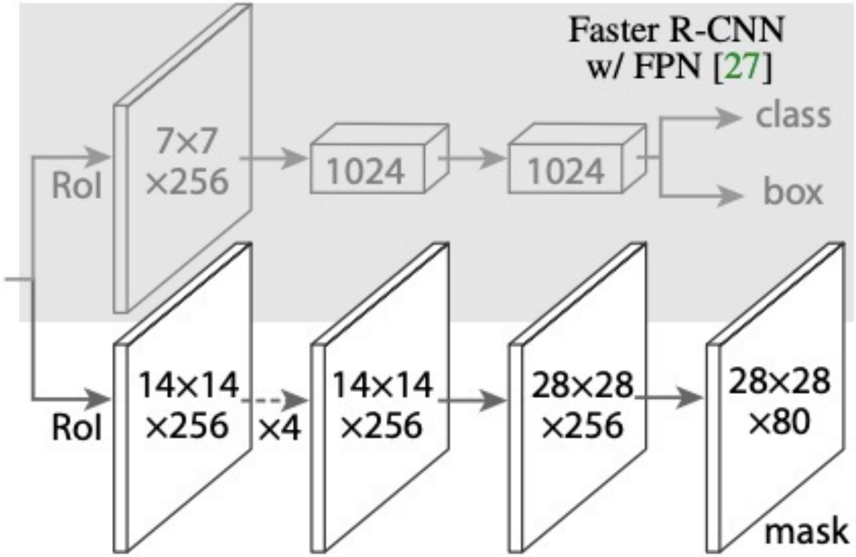

person.88
person1.00
tv.98
tv.84
person1.00
person1.00
bottle.97
wine glass.99
dining table.95
wine glass1.00
wine glass1.00

Mask Performance

	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

(c) **RoIAlign** (ResNet-50-C4): Mask results with various RoI layers. Our RoIAlign layer improves AP by $\sim$3 points and AP_{75} by $\sim$5 points. Using proper alignment is the only factor that contributes to the large gap between RoI layers.

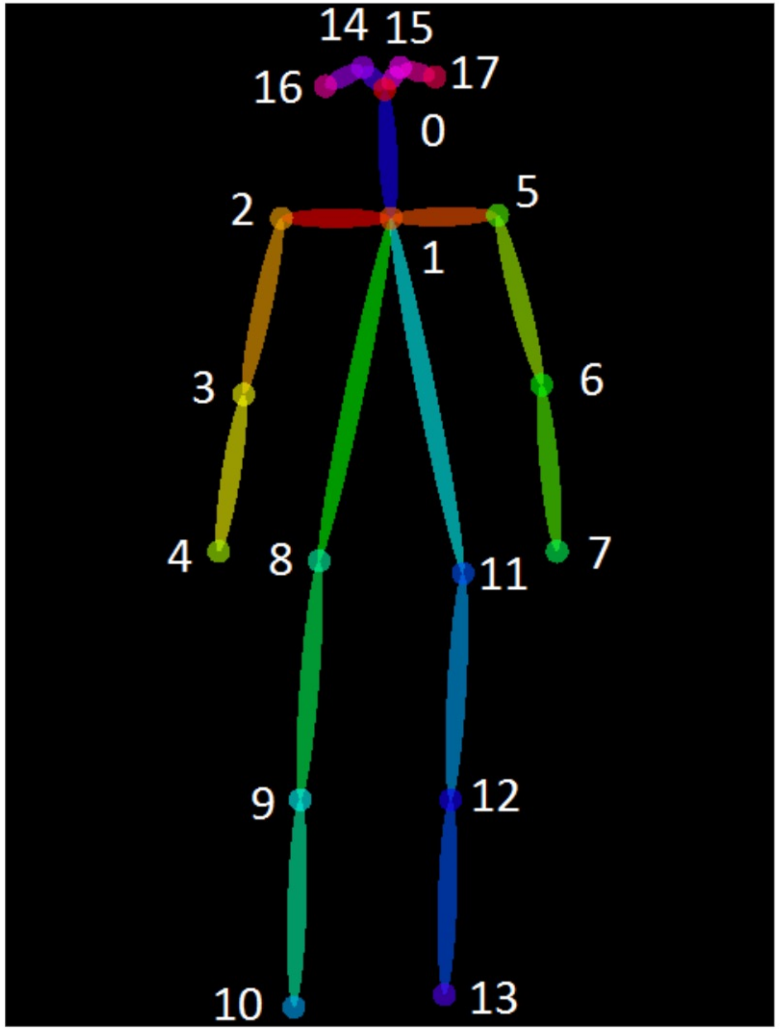
Mask Performance



	mask branch	AP	AP_{50}	AP_{75}
MLP	fc: $1024 \rightarrow 1024 \rightarrow 80 \cdot 28^2$	31.5	53.7	32.8
MLP	fc: $1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 80 \cdot 28^2$	31.5	54.0	32.6
FCN	conv: $256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 80$	**33.6**	**55.2**	**35.3**

(e) **Mask Branch** (ResNet-50-FPN): Fully convolutional networks (FCN) *vs*. multi-layer perceptrons (MLP, fully-connected) for mask prediction. FCNs improve results as they take advantage of explicitly encoding spatial layout.

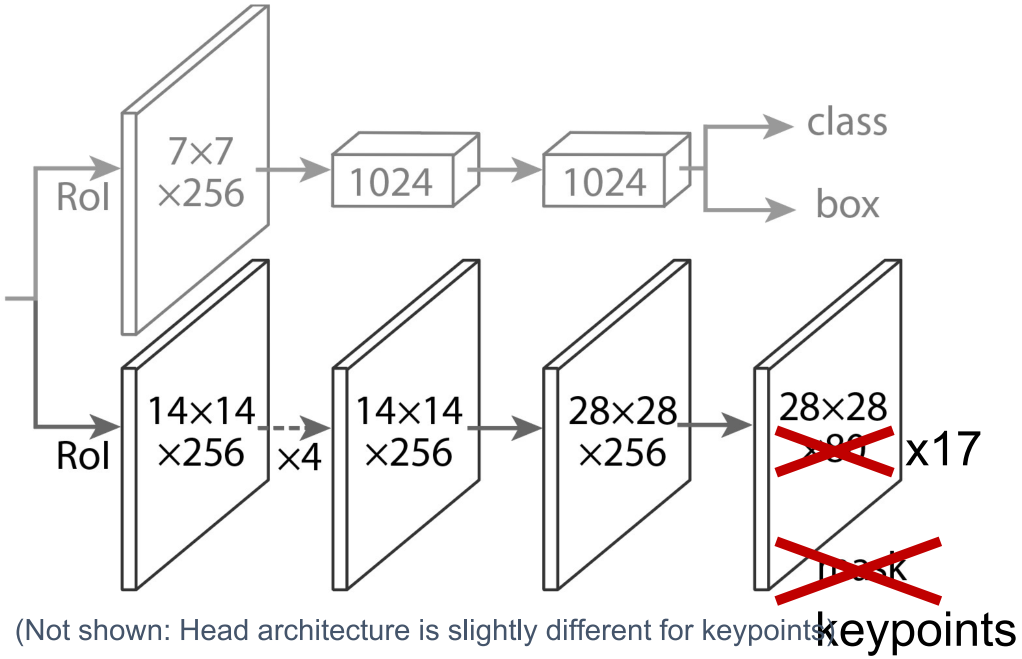
Human Pose Estimation



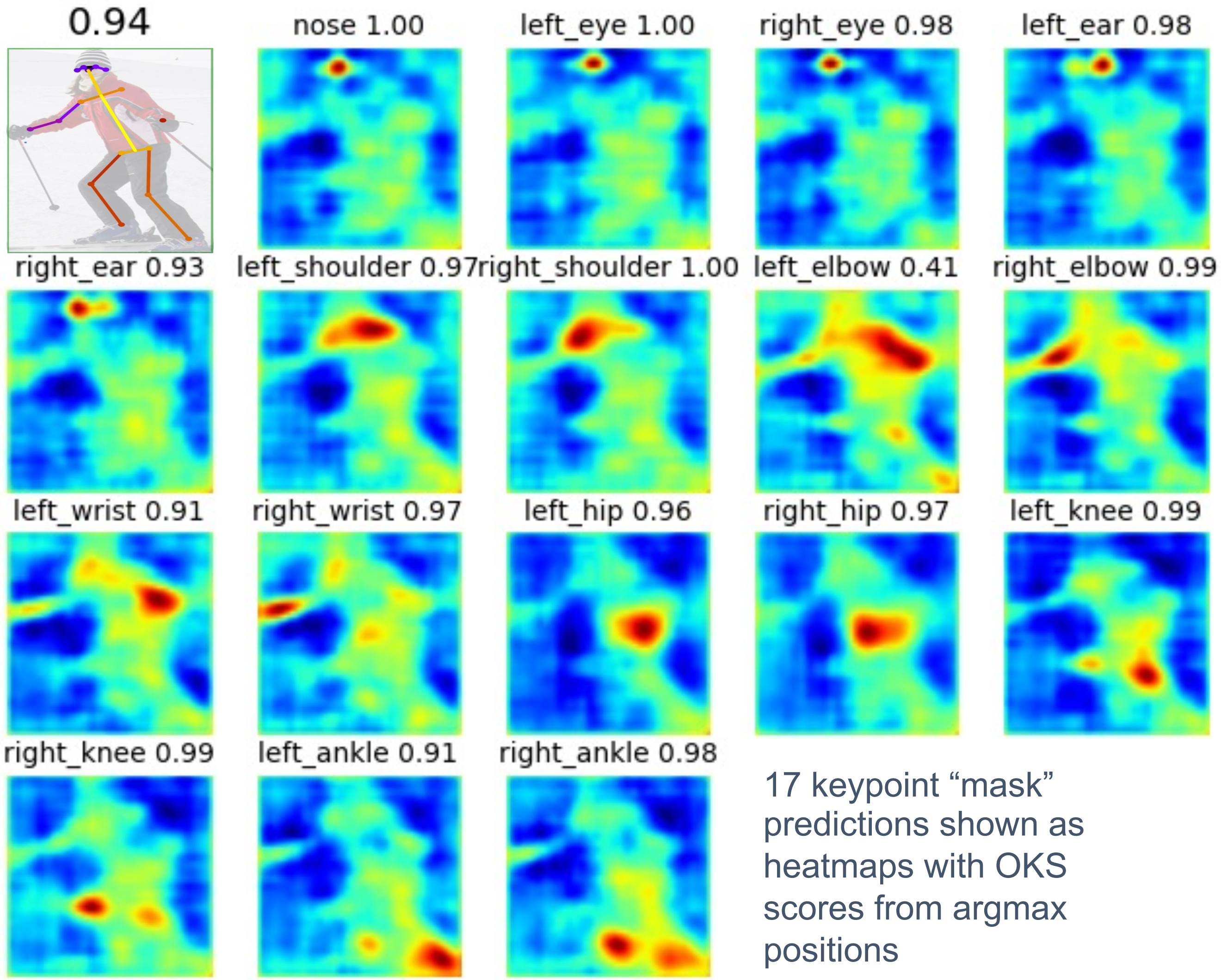
Human Pose Estimation

Human Pose GT generation

Pose Head



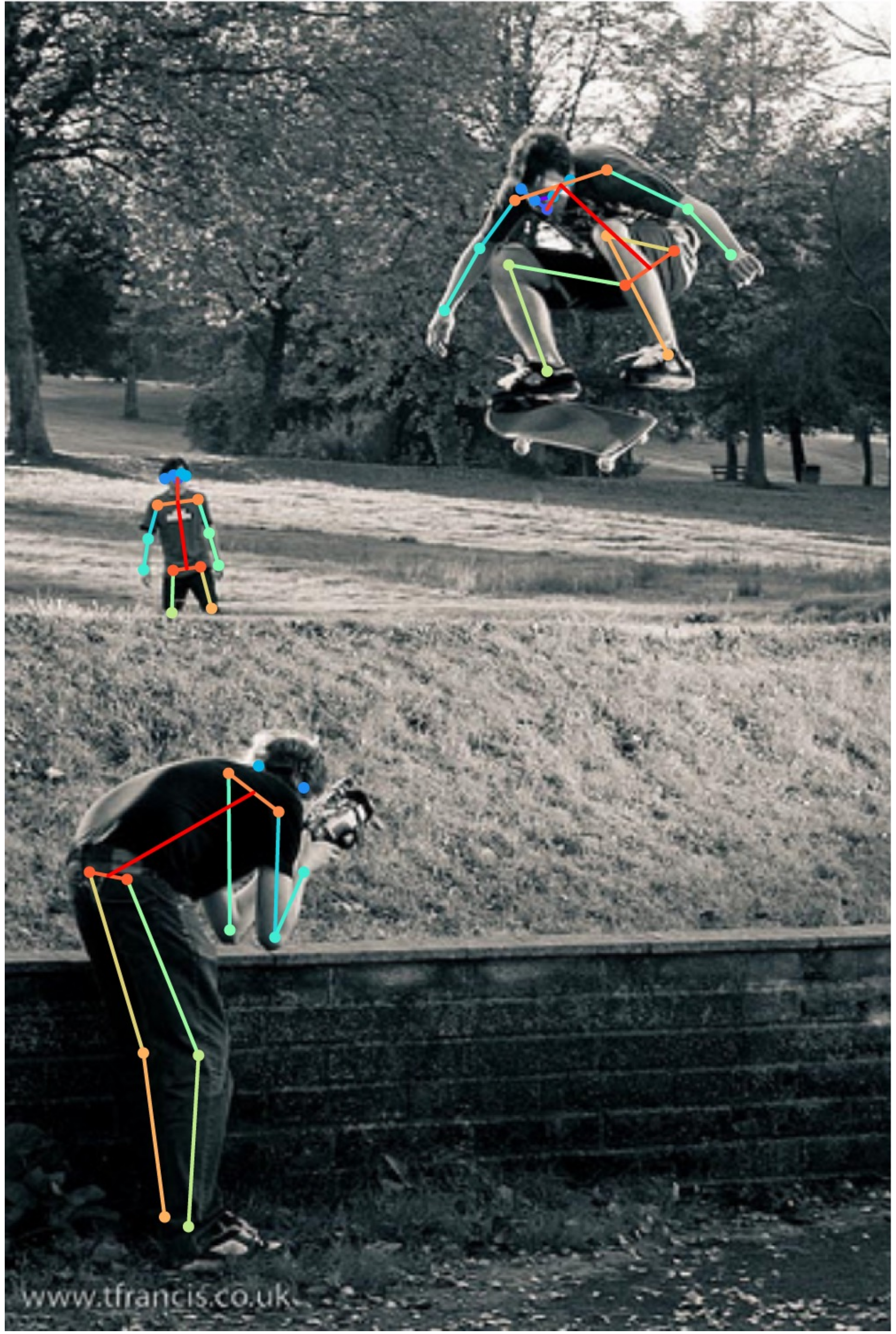
(Not shown: Head architecture is slightly different for keypoints)



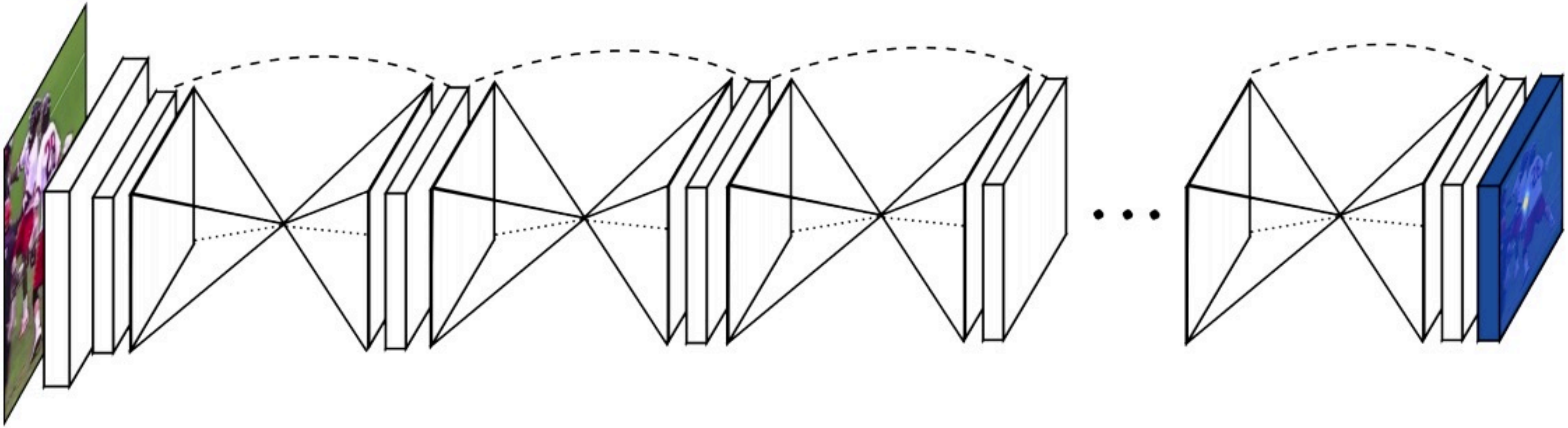
17 keypoint “mask” predictions shown as heatmaps with OKS scores from argmax positions

- Add keypoint head (28x28x17)
- Predict one “mask” for each keypoint
- Softmax over spatial locations (encodes one keypoint per mask “prior”)

Pose Head

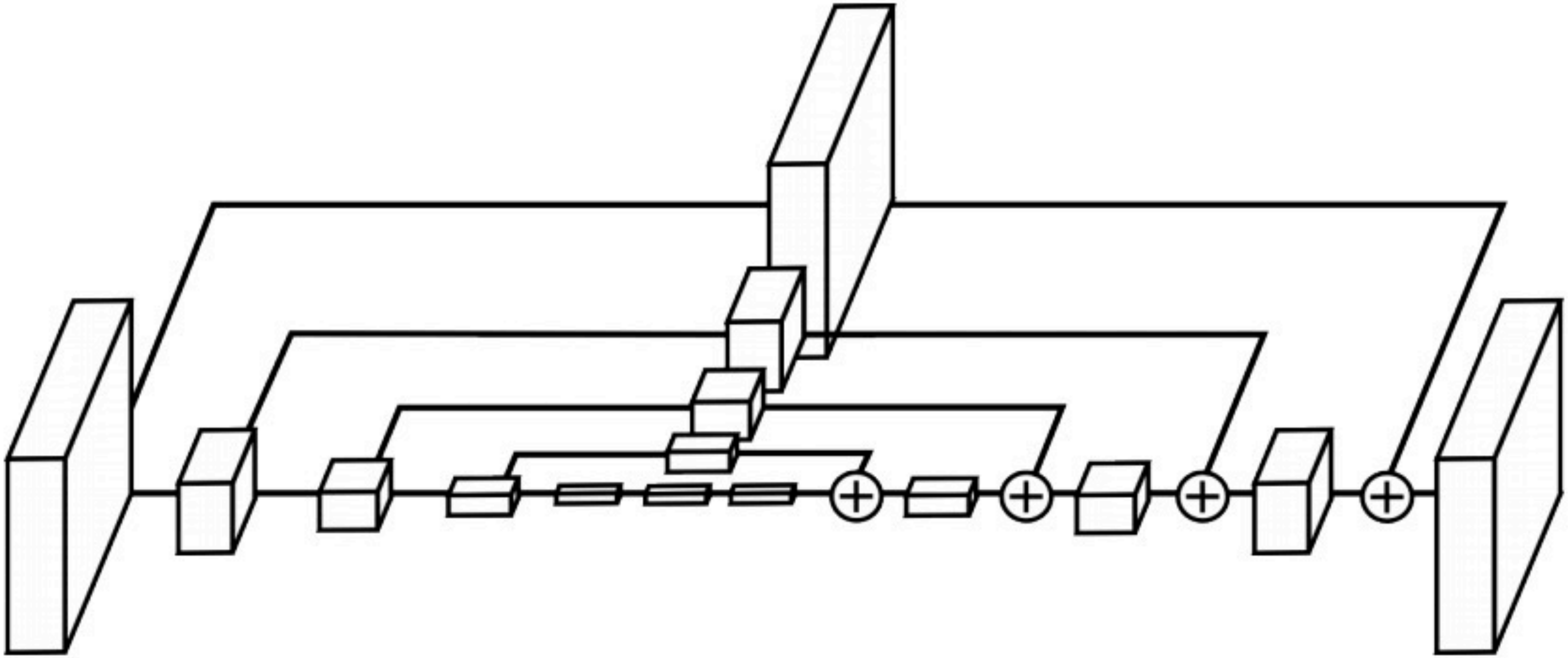


Hourglass Network for Human Pose

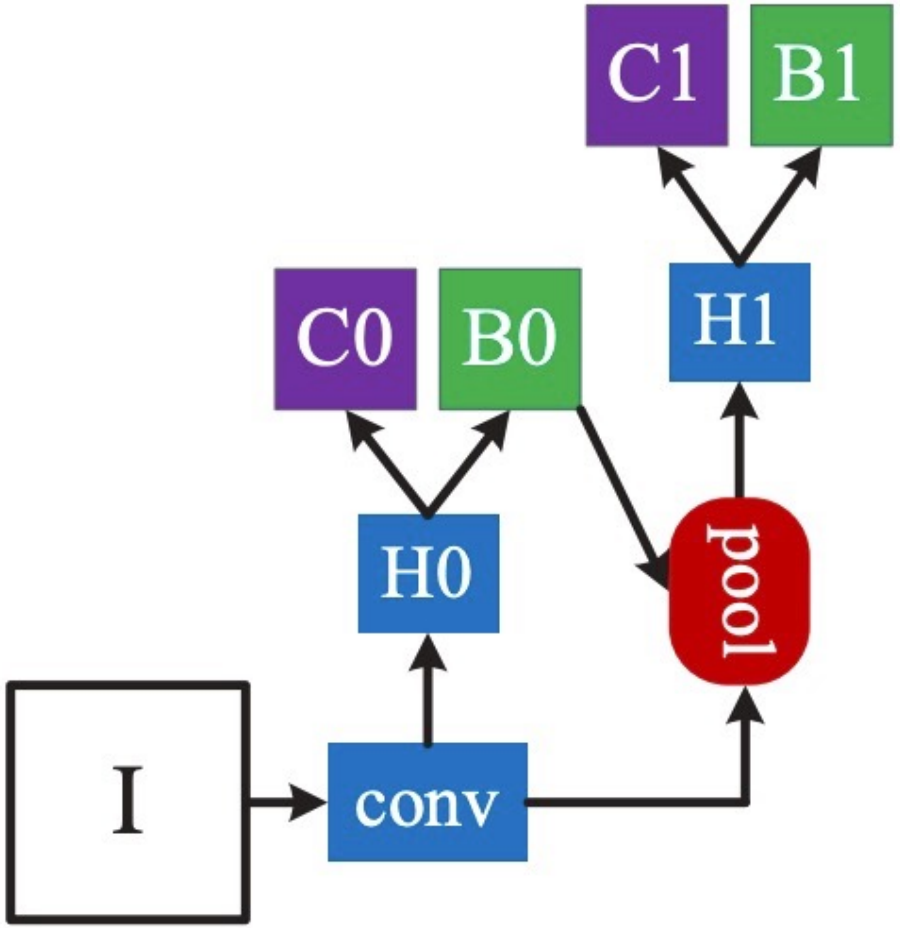


Newell et al., 2016

Hourglass Network for Human Pose



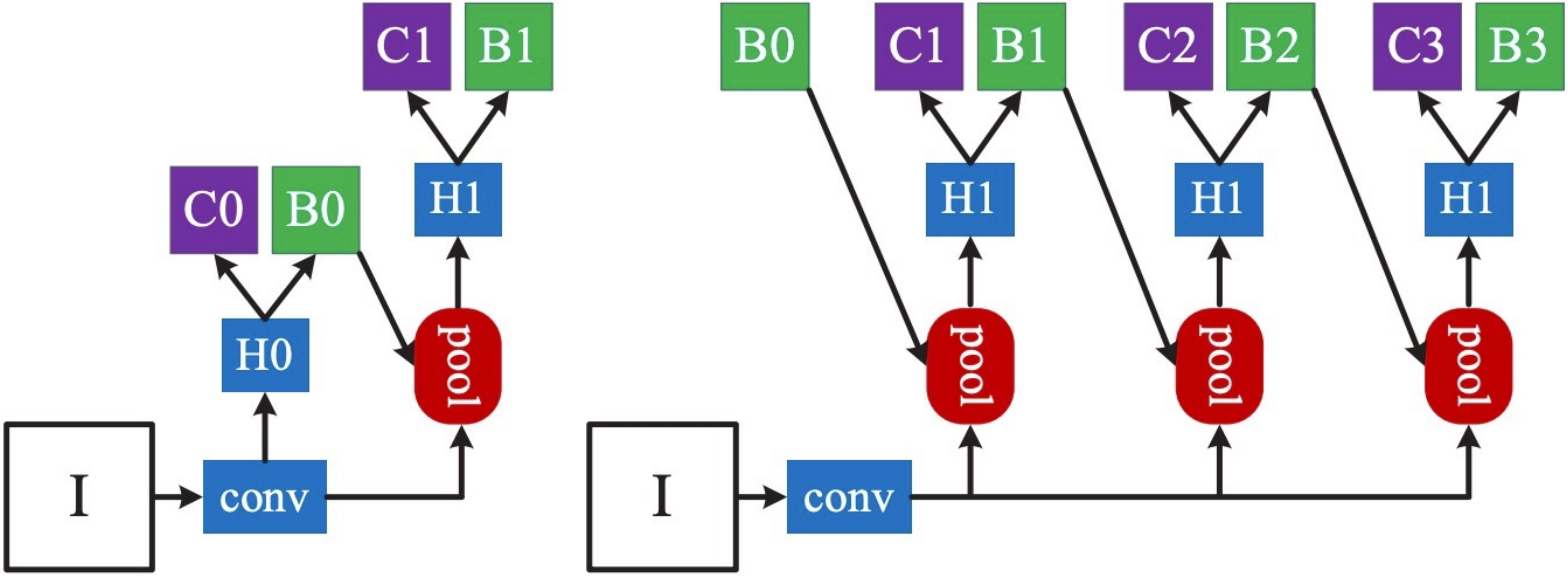
Cascade R-CNN



(a) Faster R-CNN

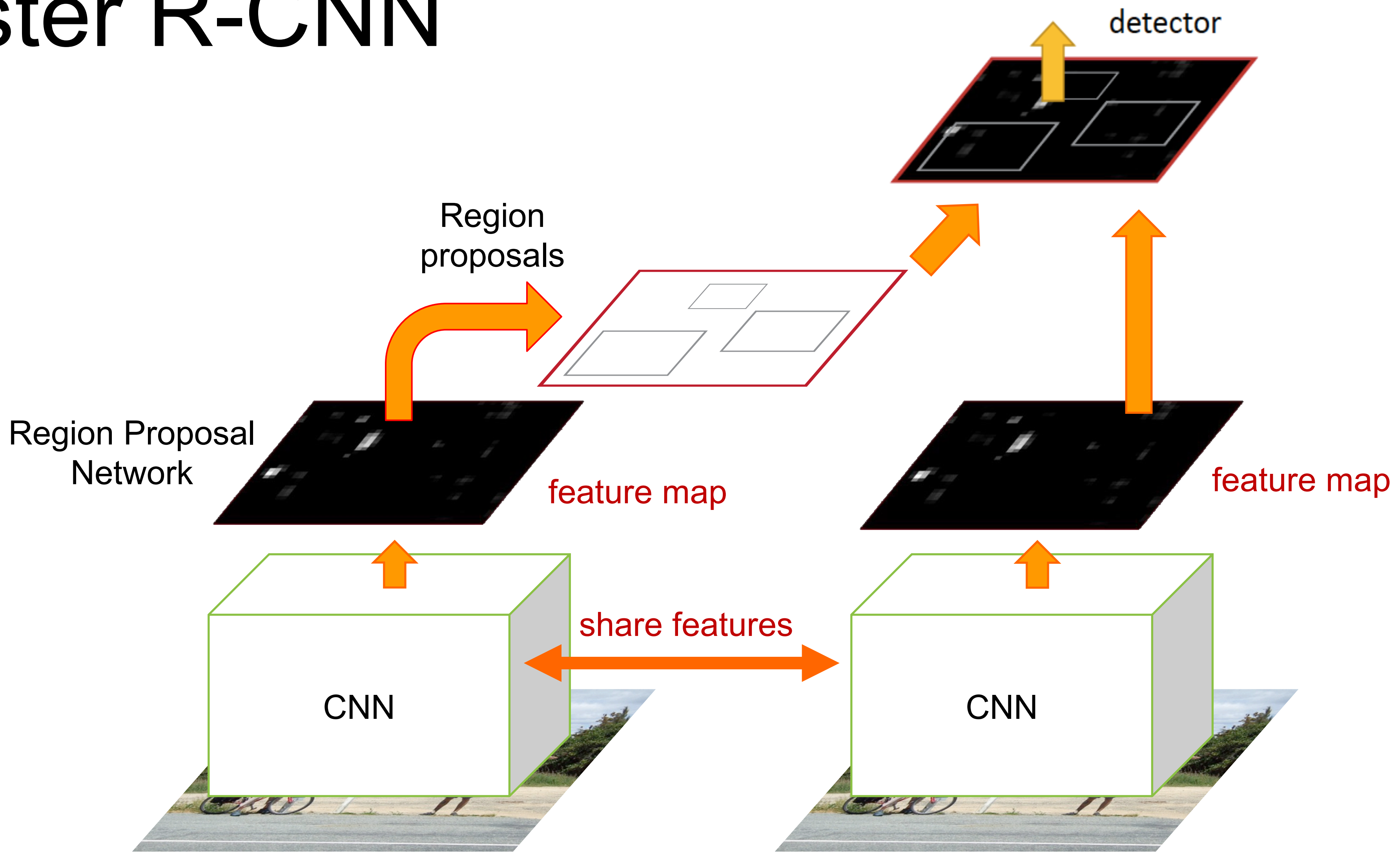
Cai et al., 2017

Cascade R-CNN



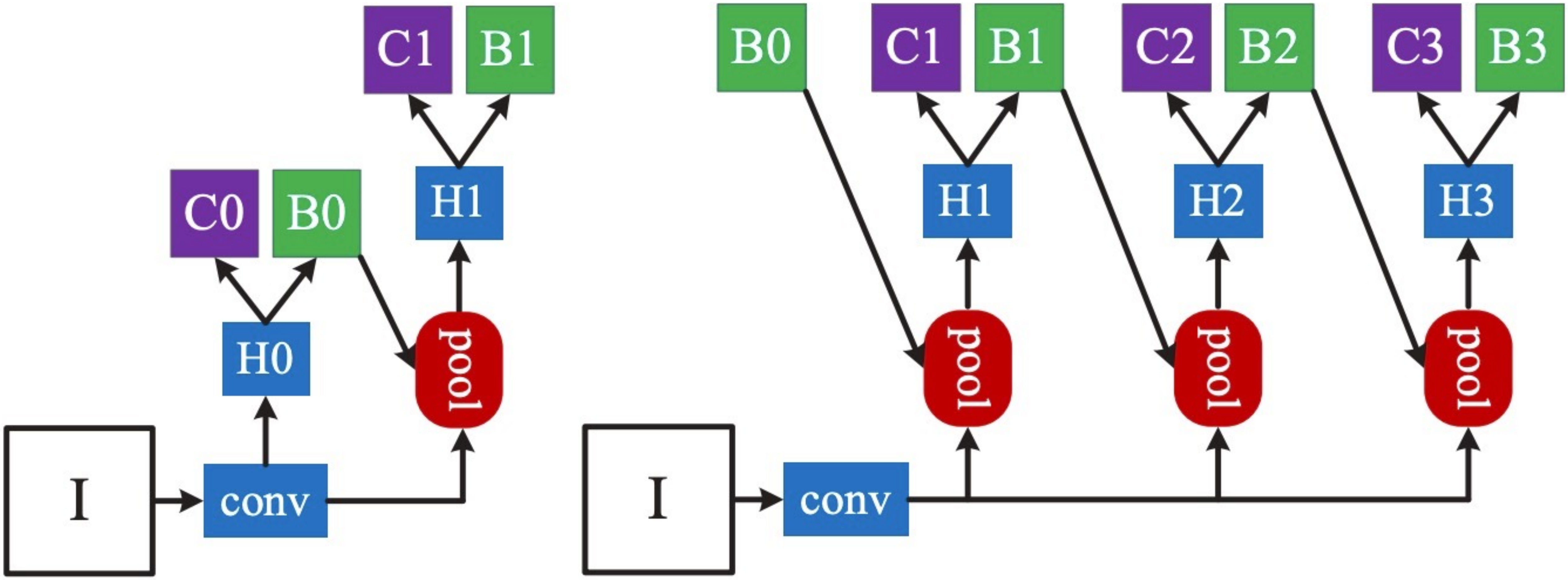
(a) Faster R-CNN (b) Iterative BBox at inference

Faster R-CNN



Ren, et al., NIPS 2015

Cascade R-CNN



(a) Faster R-CNN (d) Cascade R-CNN

Cascade R-CNN

	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
FPN+ baseline	34.9	57.0	51.9	43.6	29.7	7.1
Iterative BBox	35.4	57.2	52.1	44.2	30.4	8.1
Integral Loss	35.4	57.3	52.5	44.4	29.9	6.9
Cascade R-CNN	**38.9**	**57.8**	**53.4**	**46.9**	**35.8**	**15.8**