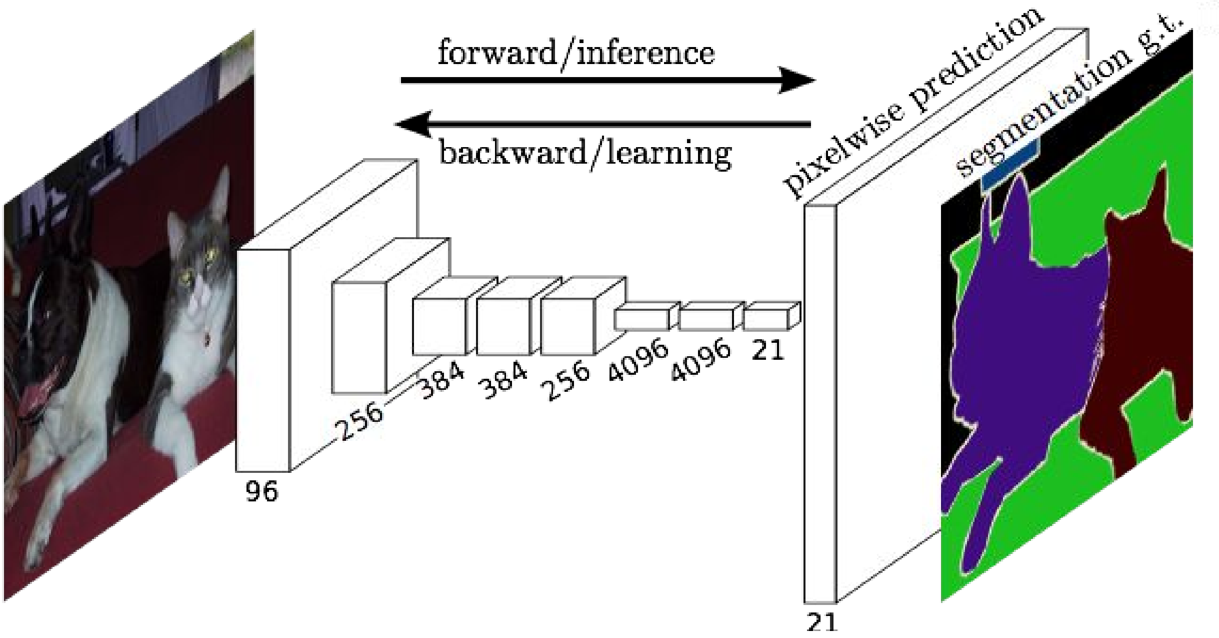
Fully Convolutional Networks for Semantic Segmentation



Jonathan Long* Evan Shelhamer* Trevor Darrell
UC Berkeley

Slides are additions from the author’s slides.

Contents

- CNN reviews
- Contributions
- Technical details

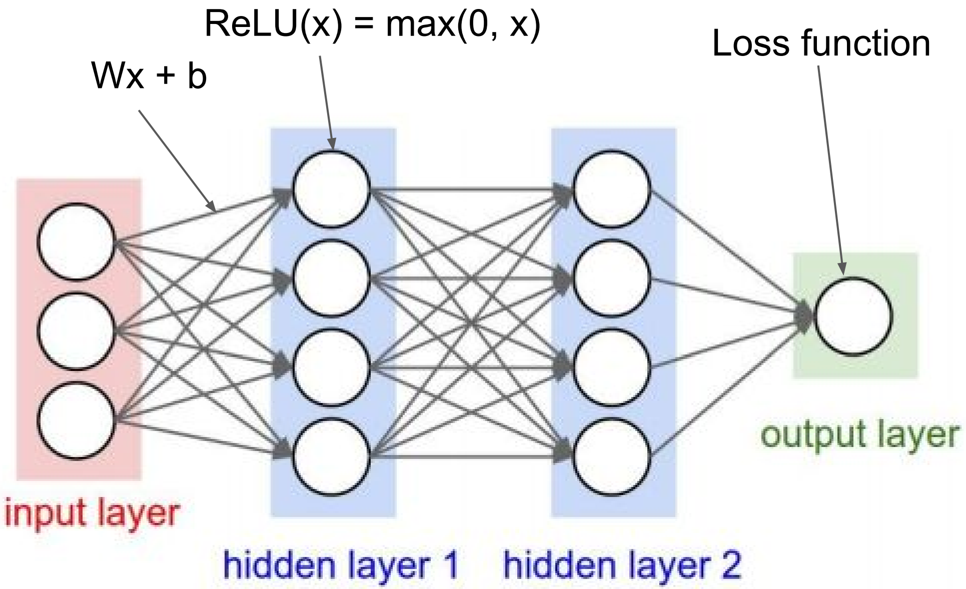
CNN reviews

A machine learning algorithm usually corresponds to a combinations of the following 3 elements:

(either explicitly specified or implicit)

- The choice of a specific function family: F (often a parameterized family).
- A way to evaluate the quality of a function $f \in F$ (typically using a cost (or loss) function L measuring how wrongly *f* predicts).
- A way to search for the «best» function $f \in F$ (typically an optimization of function parameters to minimize the overall loss over the training set).

Function family F: Multilayer Neural Networks



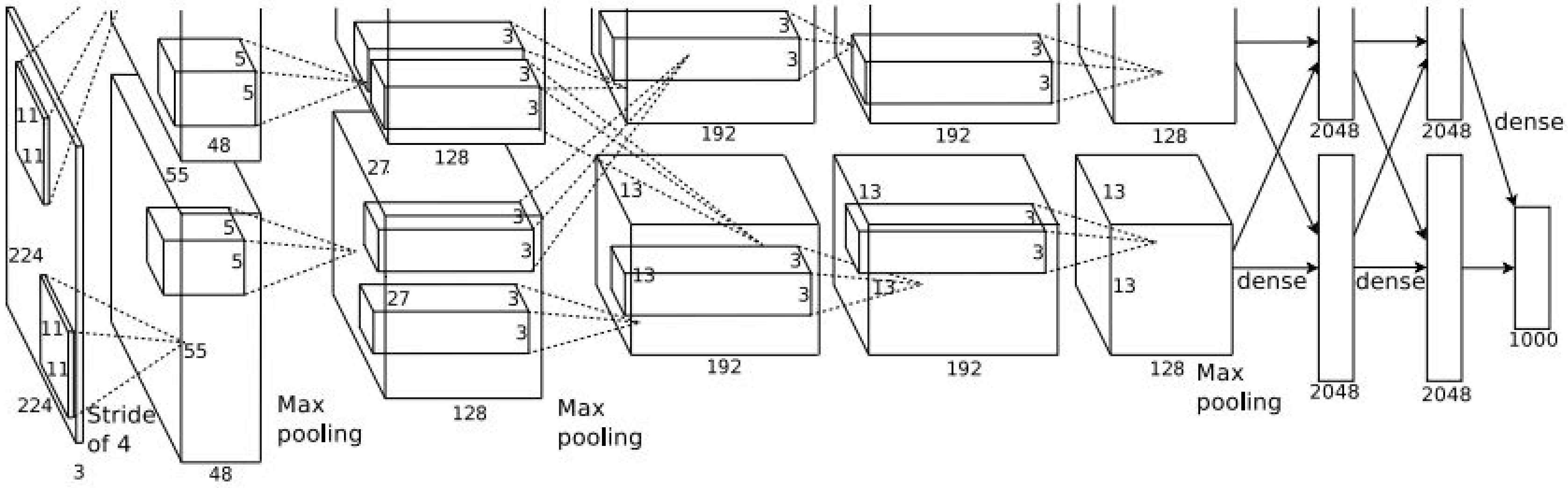
Loss function L

- For classification problem, usually cross-entropy loss function

$$J(\theta) = -\left[\sum_{i=1}^{m}\sum_{k=1}^{K} 1\left\{y^{(i)} = k\right\} \log P(y^{(i)} = k | x^{(i)}; \theta)\right]$$

Modern CNNs

- AlexNet (Krizhevsky *et. al.*)



Architecture terminology

- Convolution layer: C(k,n,s), pooling layer: P(k,s), normalization layer: N, rectified linear unit: RL
- AlexNet: **C**(11x11x3, 96, 4) - **RL** - **N** - **P**(3, 2) - **C**(5x5x48, 256, 1) - **RL** - **N** - **P**(3, 2) - **C**(3x3x256, 384, 1) - **RL** - **C**(3x3x384, 384, 1) - **RL** - **C**(3x3x384, 256, 1) - **RL** - **P**(3, 2) - **FC**(4096) - **D**(0.5) - **FC**(4096) - **D**(0.5) - **FC**(1000)

Number of parameters in CNN

- **conv1**: 96x3x11x11 = 34 848
- **conv2**: 256x48x5x5 = 307 200
- **conv3**: 384x256x3x3 = 884 736
- **conv4**: 384x192x3x3 = 663 552
- **conv5**: 256x192x3x3 = 442 368
- **fc6**: 4096 x 9216 = 37 748 736
- **fc7**: 4096 x 4096 = 16 777 216

54 525 952 (95.2%)

- **fc8**: 101 x 4096 = 413 696
- **Total** = 57 272 352

9216=256x6x6

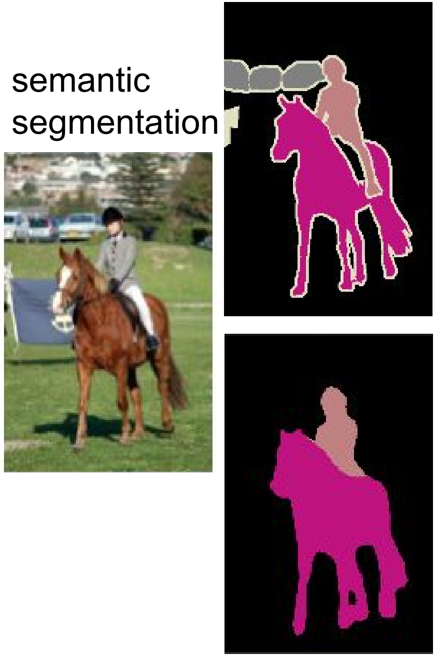
Contributions

Convolutionize a usual CNN for dense pixel predictions.

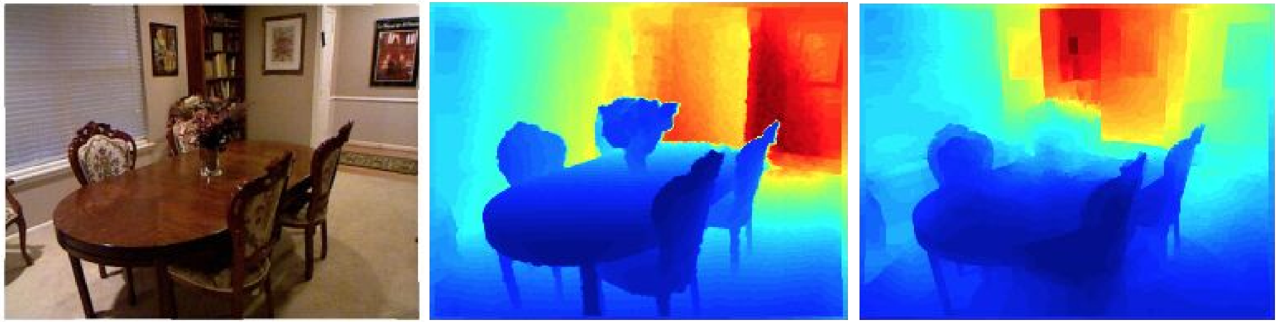
FCN

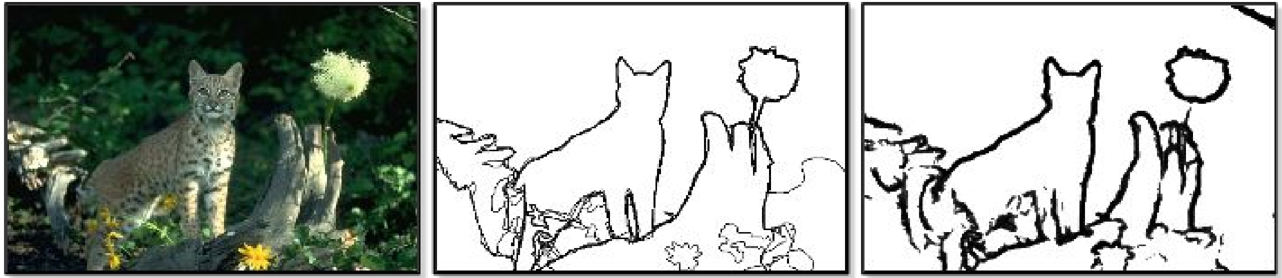
pixels in, pixels out

semantic
segmentation



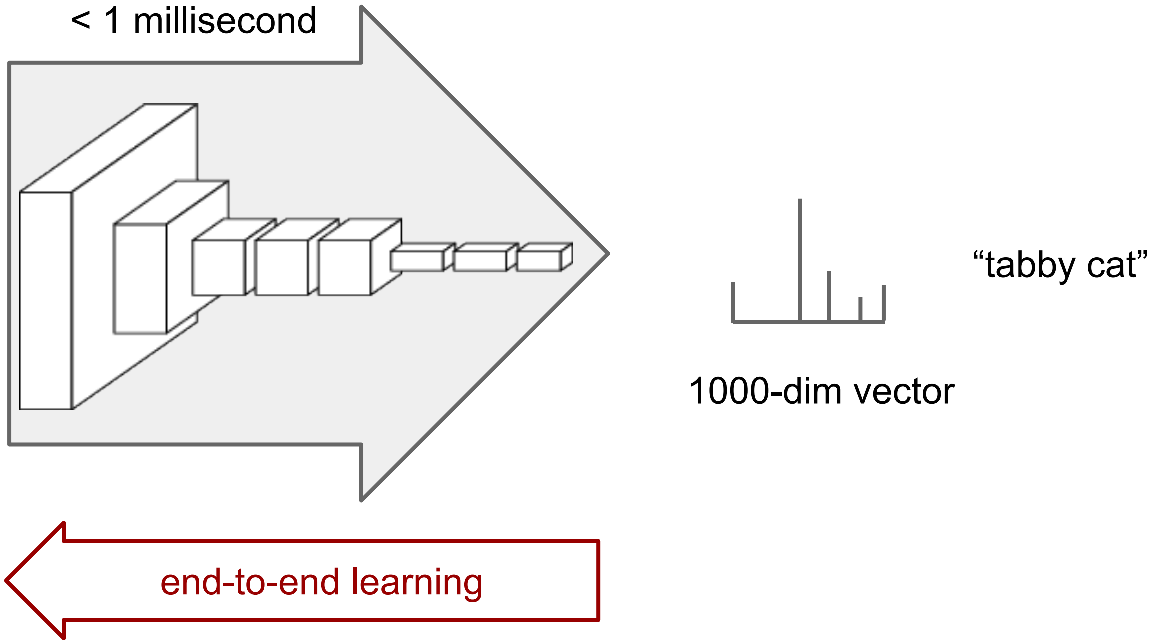
monocular depth estimation (Liu et al. 2015)



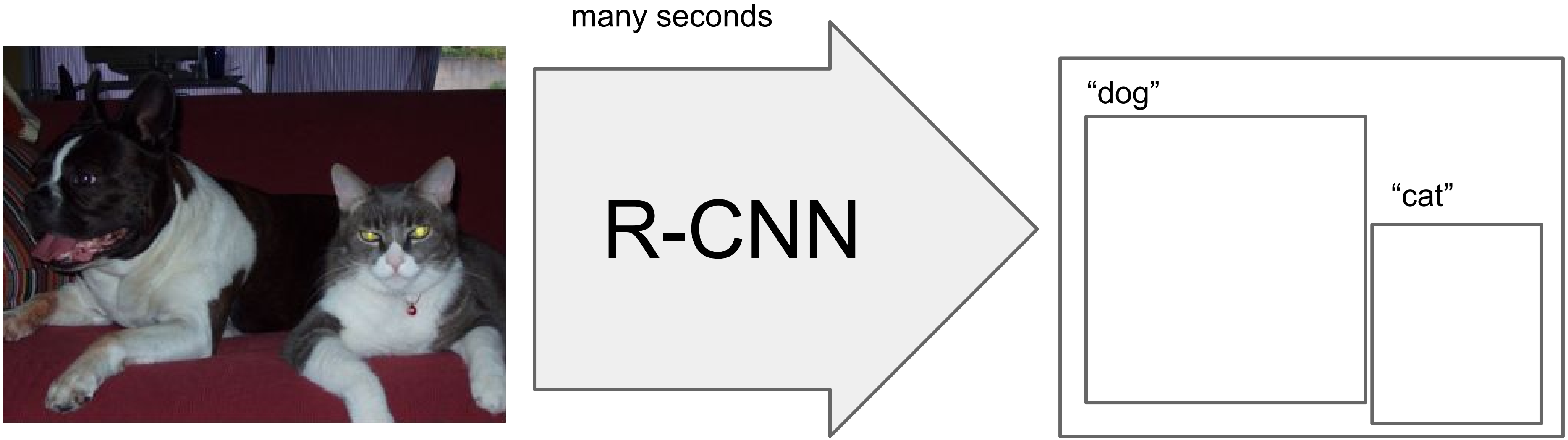


boundary prediction (Xie & Tu 2015)

convnets perform classification



R-CNN does detection



R-CNN

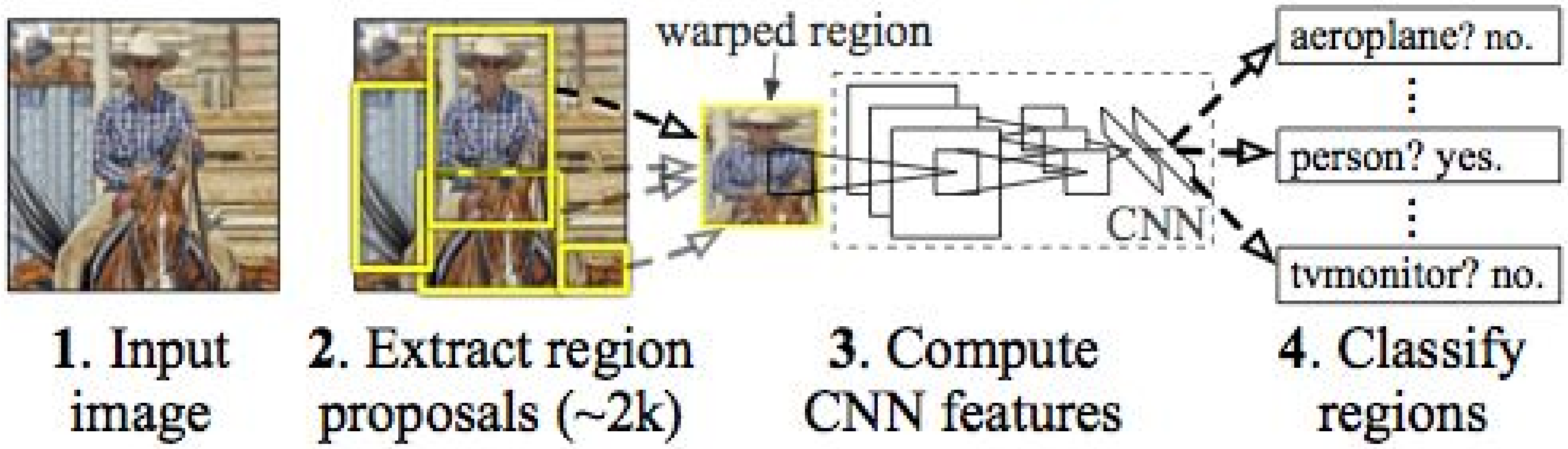
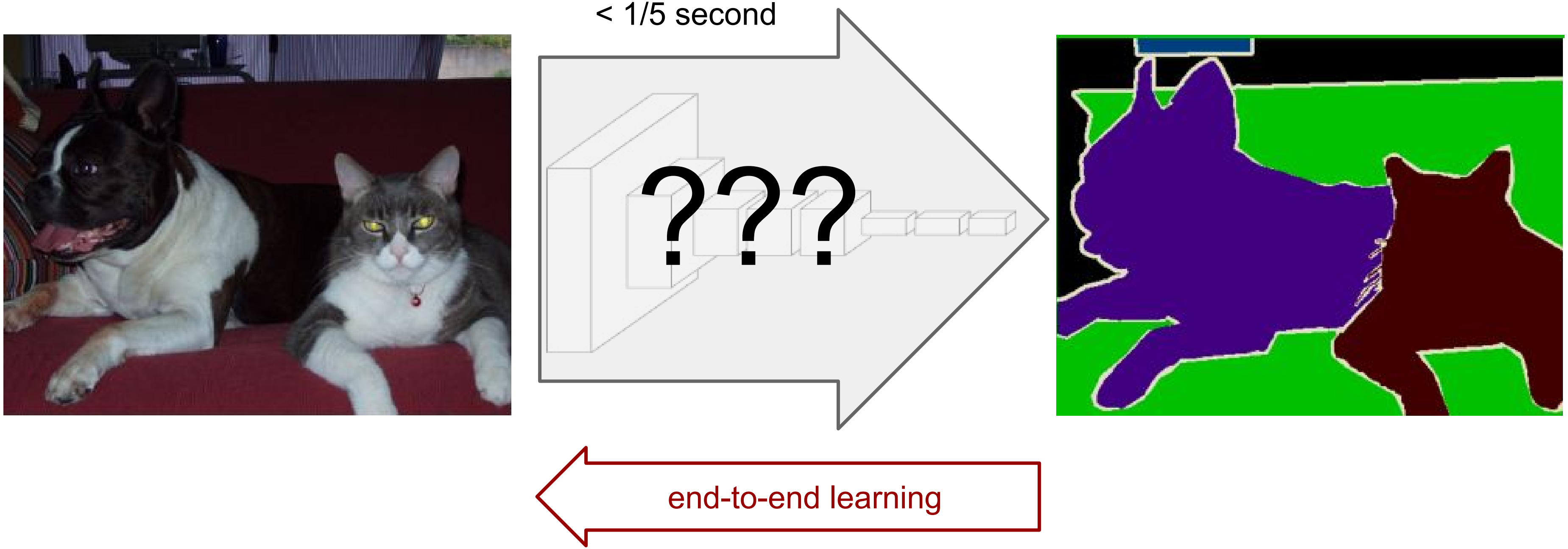
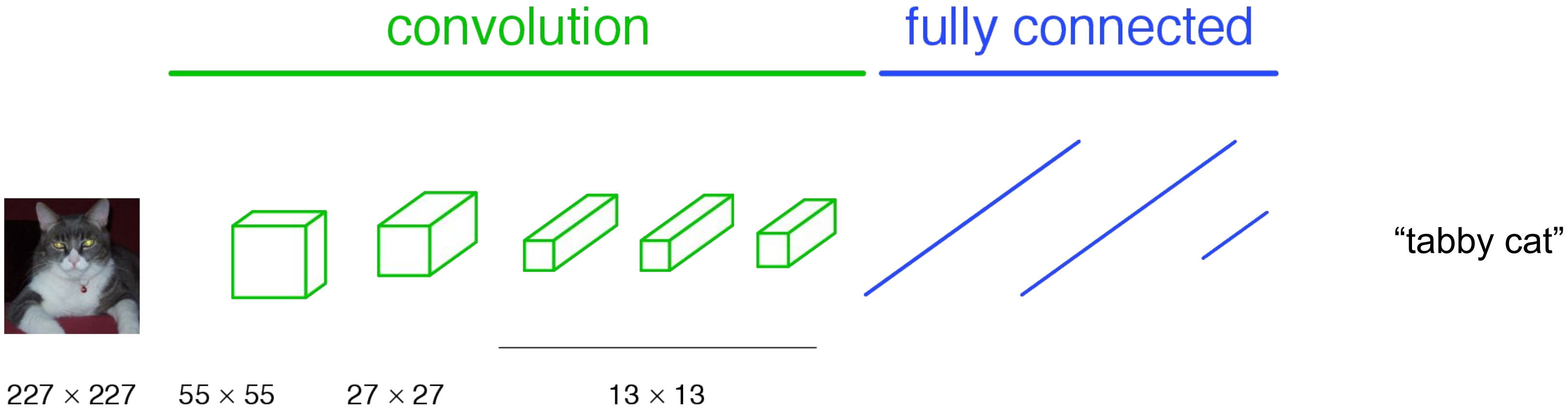


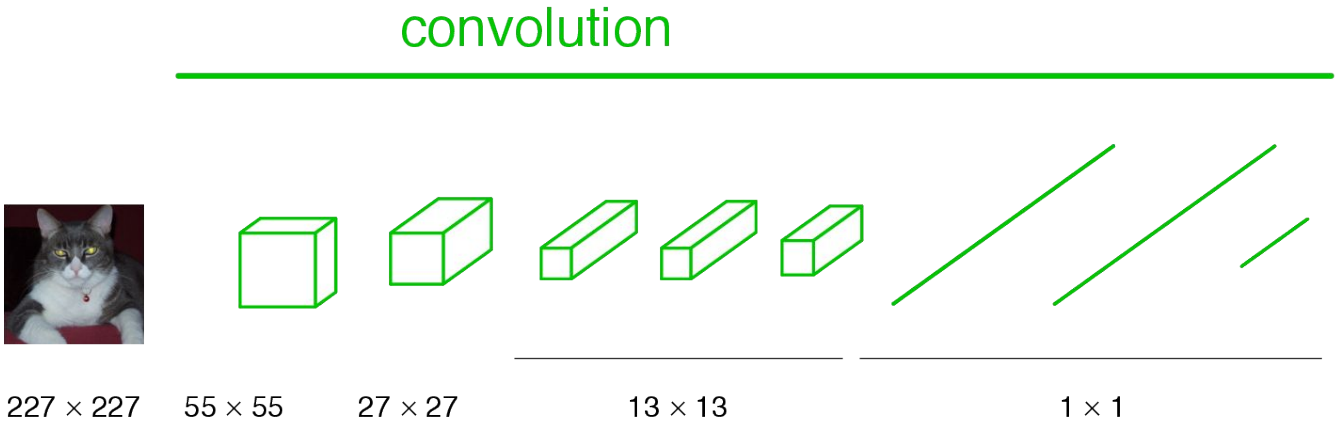
figure: Girshick et al.


< 1/5 second
???
end-to-end learning

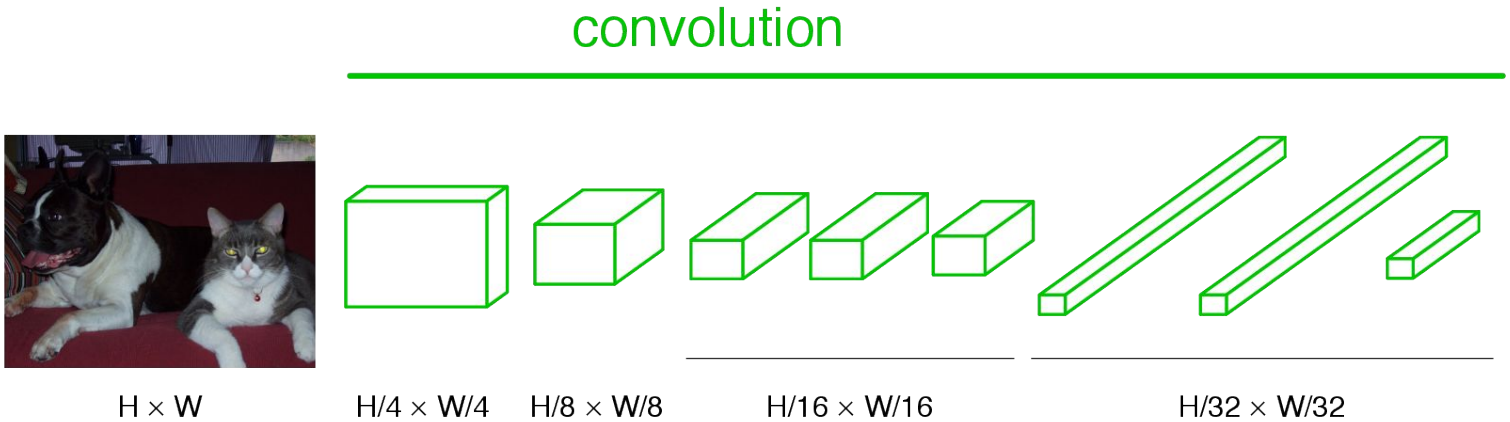
a classification network



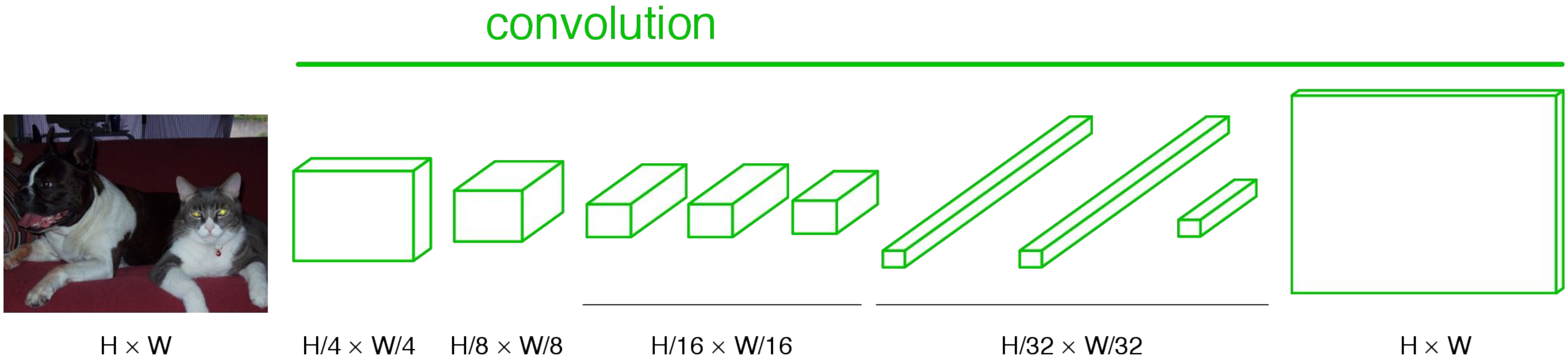
becoming fully convolutional



becoming fully convolutional

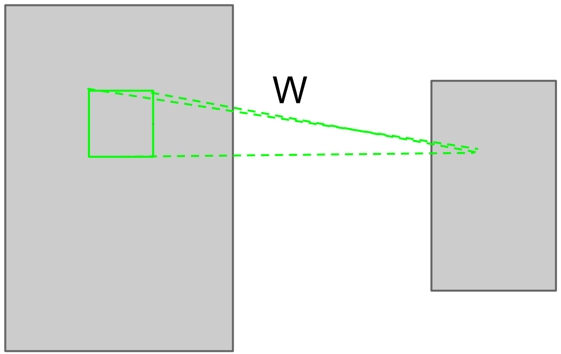


upsampling output



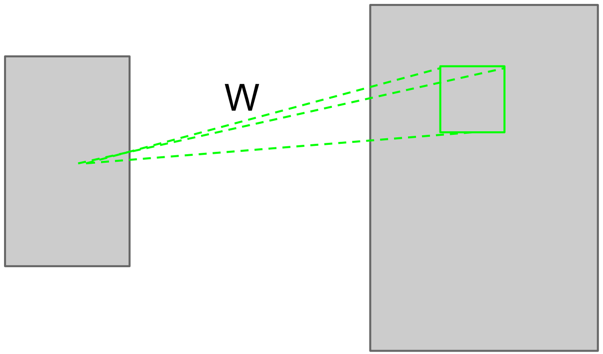
Deconvolution layer

input stride f



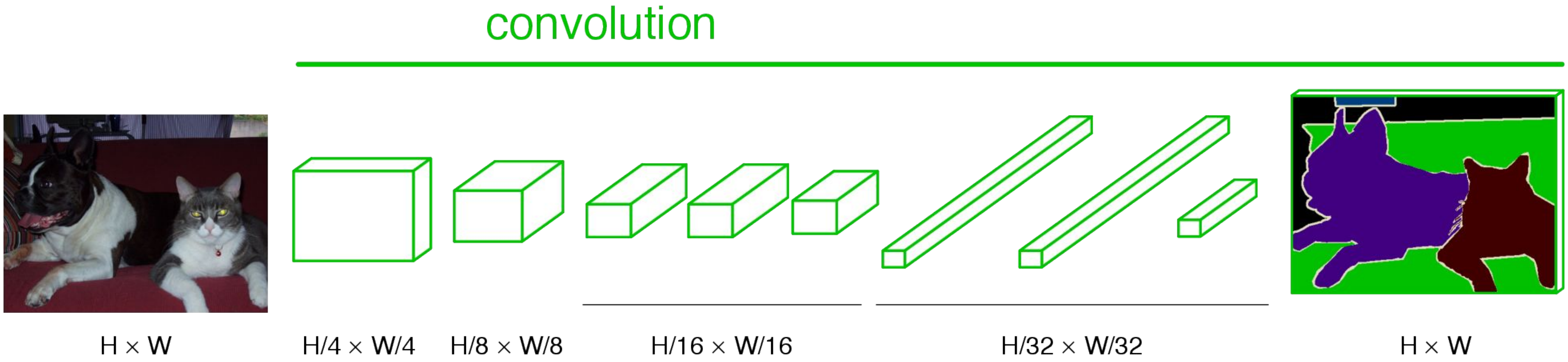
convolution

output stride f

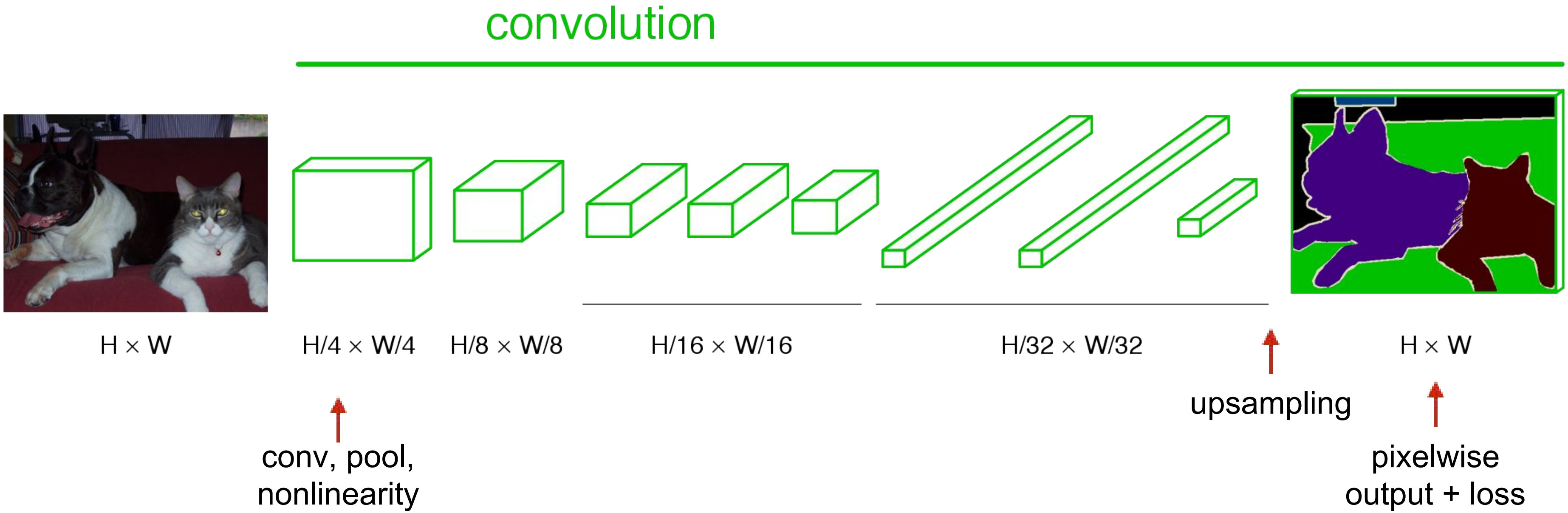


deconvolution

end-to-end, pixels-to-pixels network



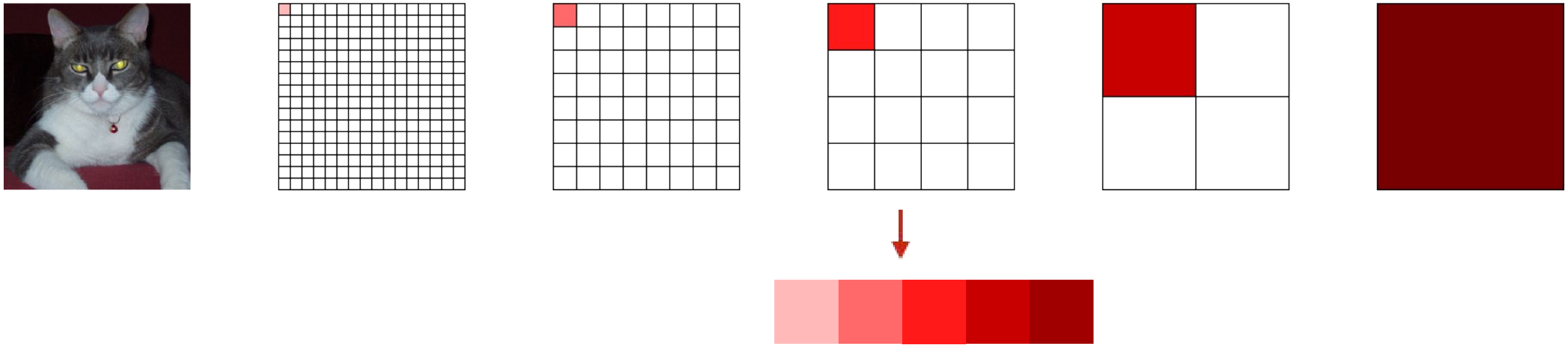
end-to-end, pixels-to-pixels network



spectrum of deep features

combine *where* (local, shallow) with *what* (global, deep)

image intermediate layers



fuse features into **deep jet**

(cf. Hariharan et al. CVPR15 “hypercolumn”)

skip layers

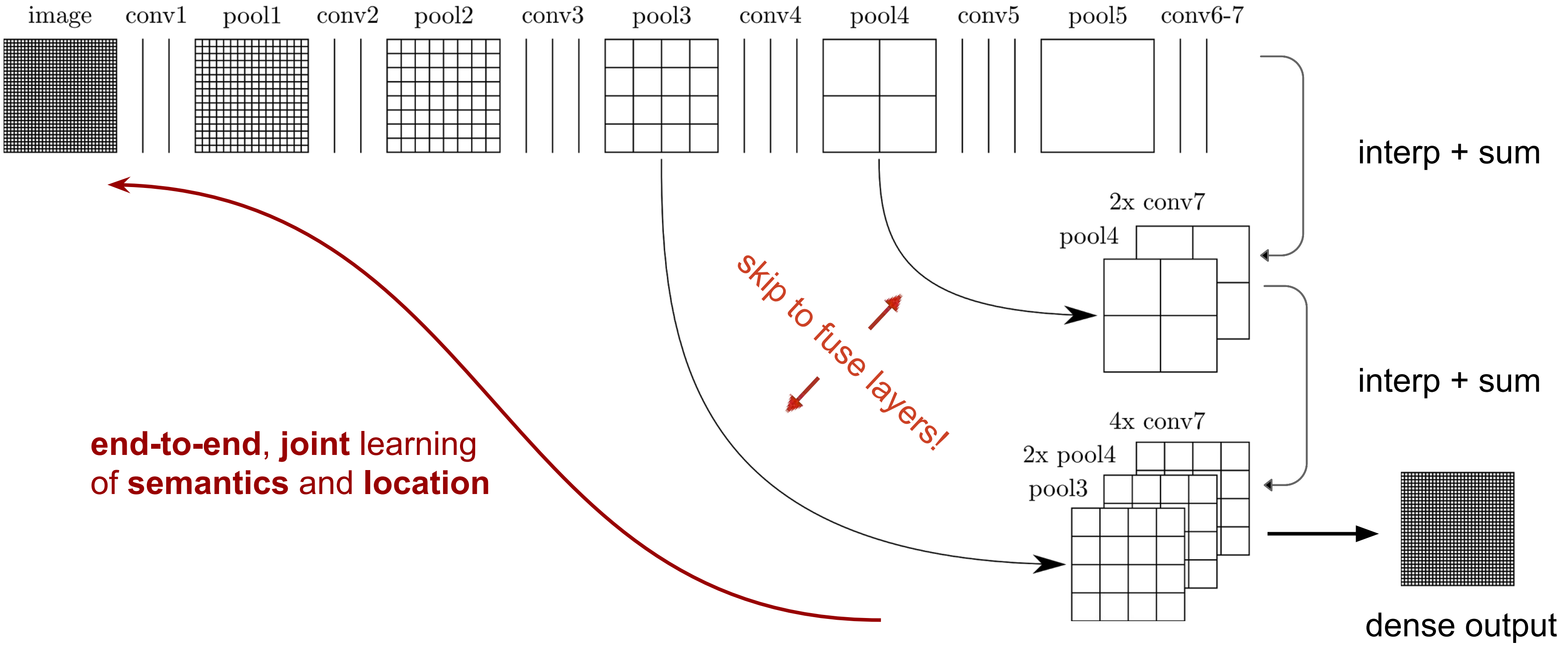
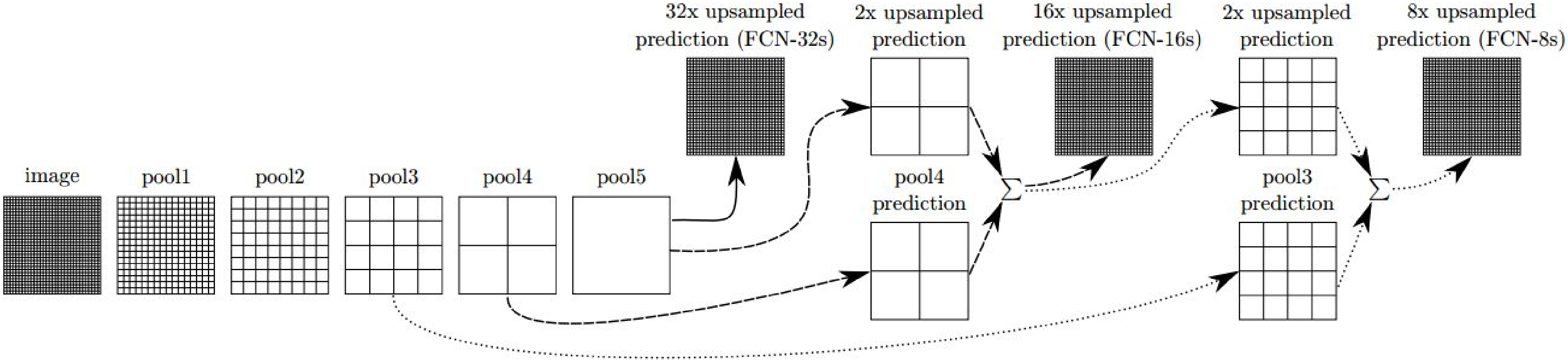


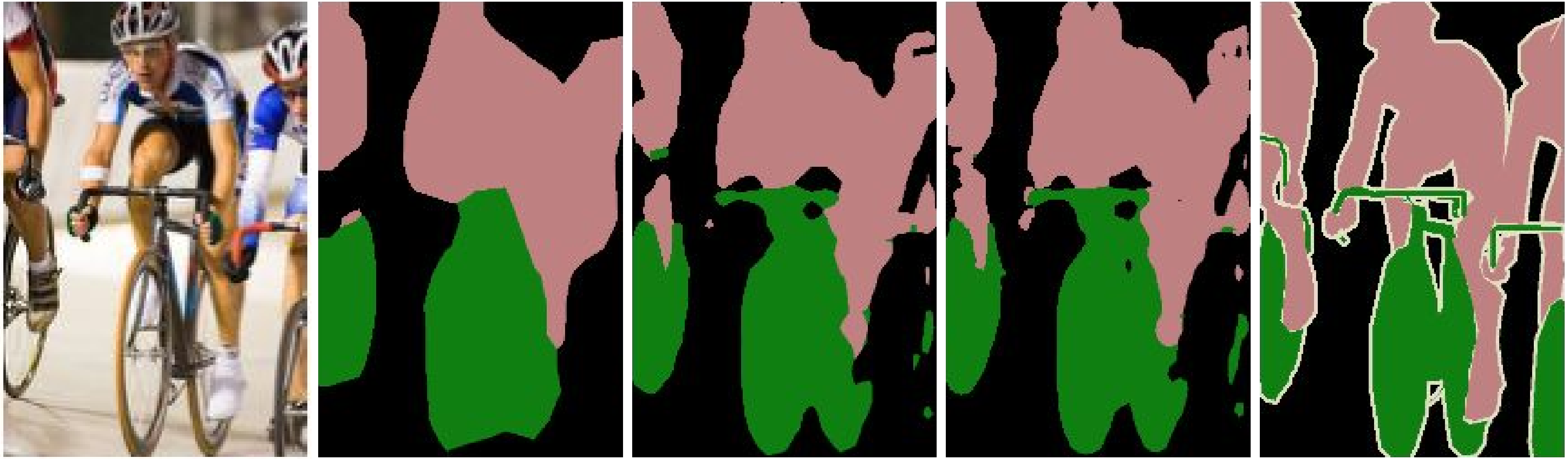
image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
interp + sum
2x conv7
pool4
skip to fuse layers!
interp + sum
4x conv7
2x pool4
pool3
end-to-end, joint learning
of semantics and location
dense output

DAG


32x upsampled prediction (FCN-32s)
2x upsampled prediction
16x upsampled prediction (FCN-16s)
2x upsampled prediction
8x upsampled prediction (FCN-8s)
image
pool1
pool2
pool3
pool4
pool5
pool4 prediction
pool3 prediction

skip layer refinement

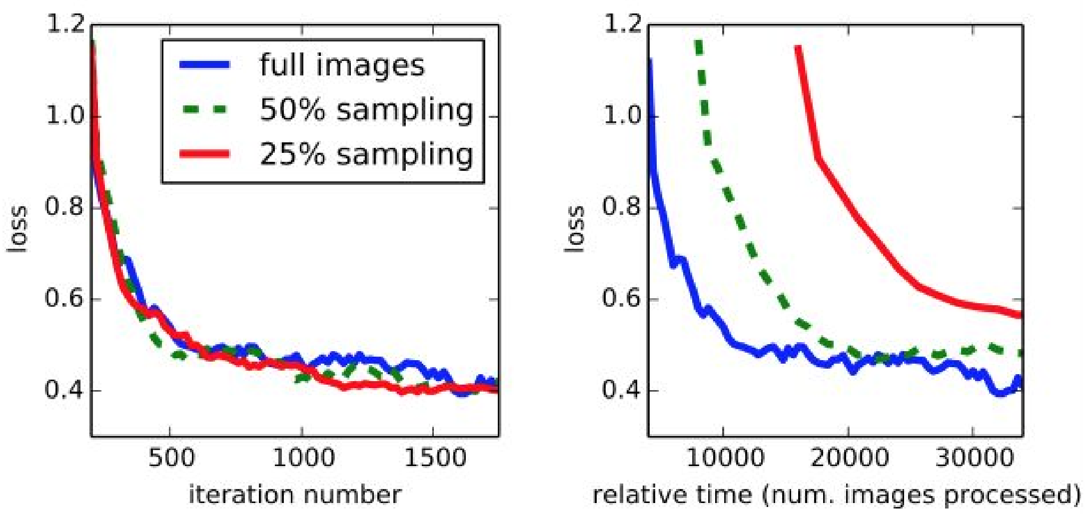
input image | stride 32 | stride 16 | stride 8 | ground truth

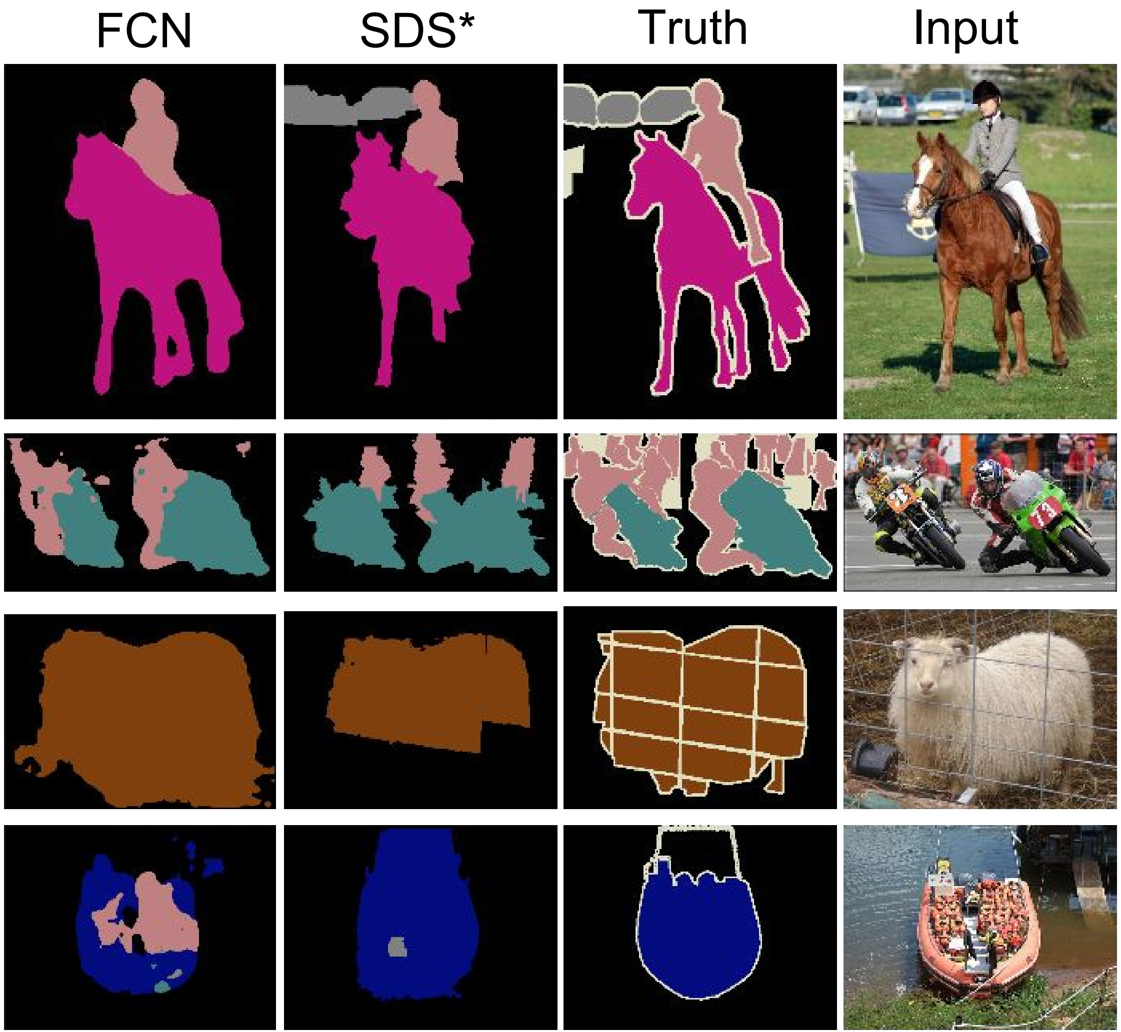


no skips | 1 skip | 2 skips

training + testing

- train full image at a time *without patch sampling*
- reshape network to take input of any size
- forward time is ~150ms for 500 x 500 x 21 output





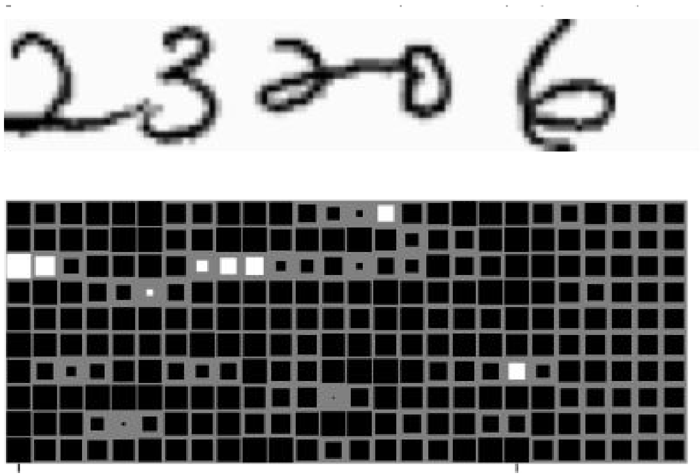
Relative to prior state-of-the-art SDS:

- 20% relative improvement for mean IoU
- 286× faster

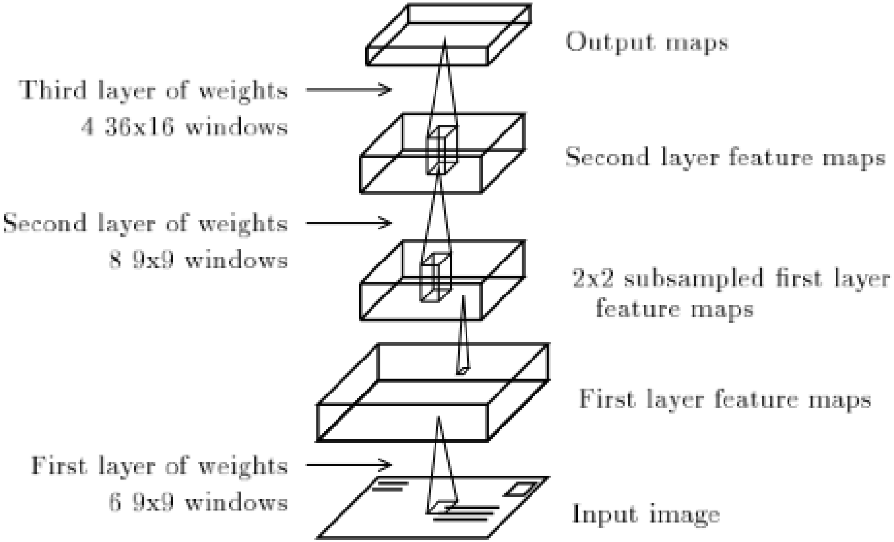
*Simultaneous Detection and Segmentation
Hariharan et al. ECCV14

past and future history of fully convolutional networks

history



Shape Displacement Network
Matan & LeCun 1992



Convolutional Locator Network
Wolf & Platt 1994

pyramids

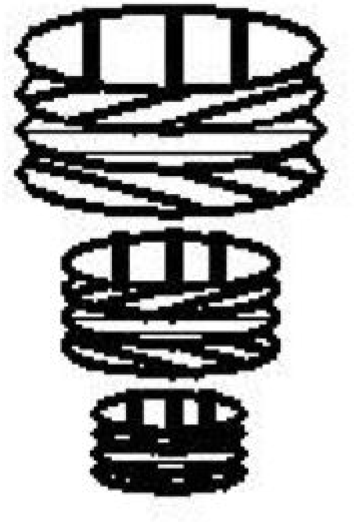
0 1 2

Scale Pyramid, *Burt & Adelson '83*

The scale pyramid is a classic multi-resolution representation.

Fusing multi-resolution network layers is a learned, nonlinear counterpart.

jets



Jet, *Koenderink & Van Doorn '87*

The local jet collects the partial derivatives at a point for a rich local description.

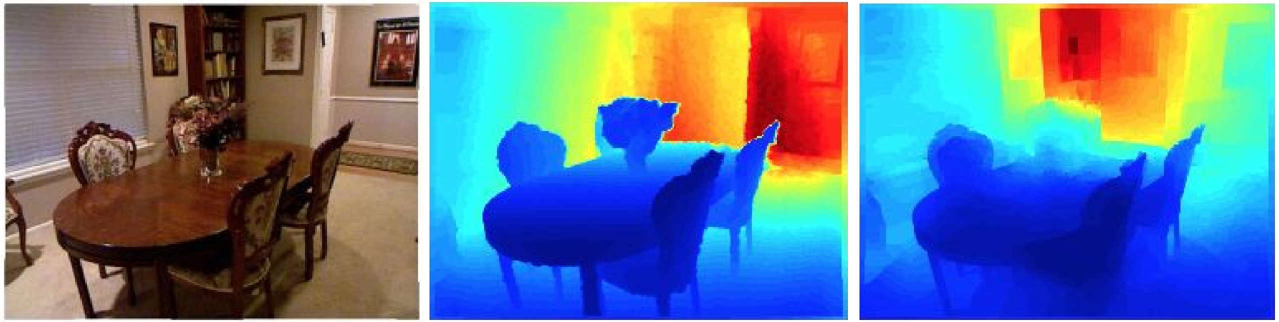
The deep jet collects layer compositions for a rich, learned description.

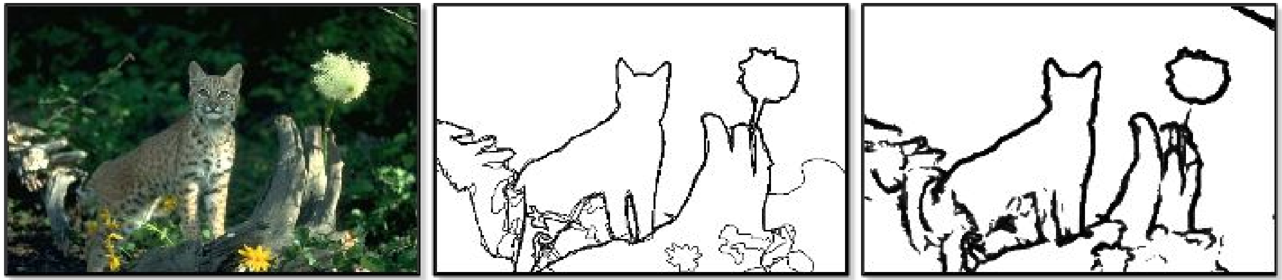
extensions

- more tasks
- random fields
- weak supervision

image to image learning

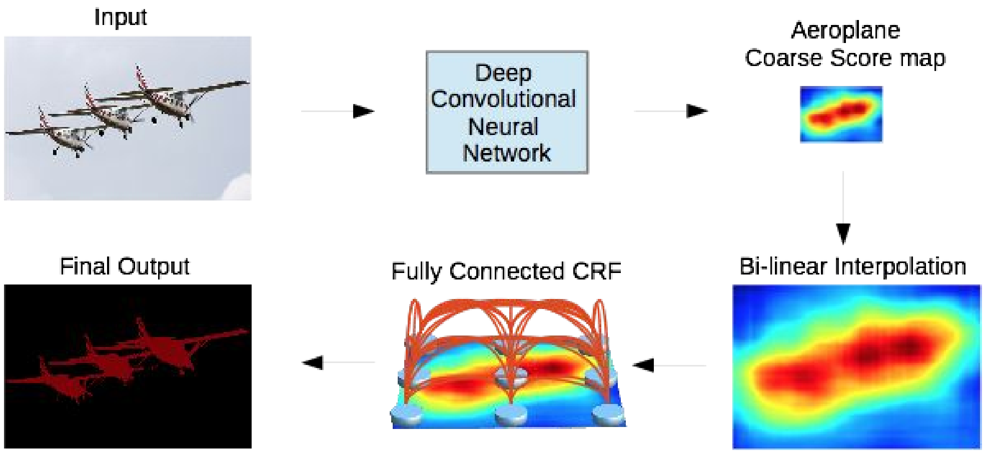
monocular depth estimation (Liu et al. 2015)

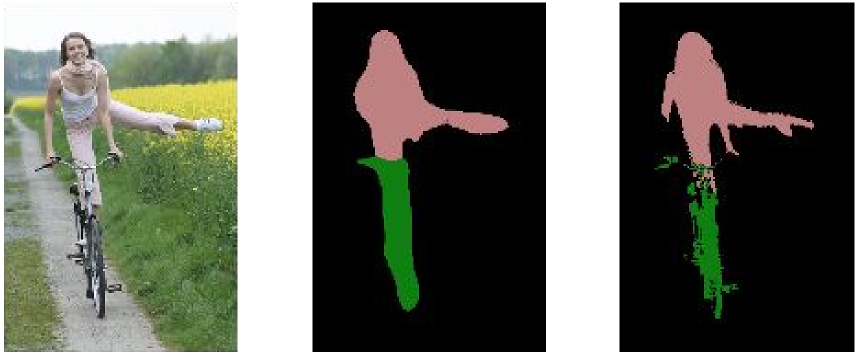




boundary prediction (Xie & Tu 2015)

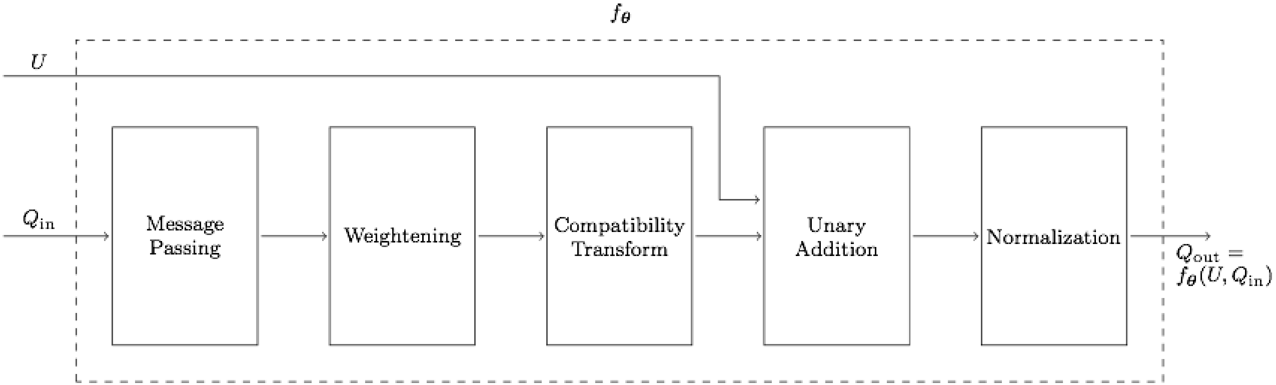
fully conv. nets + random fields





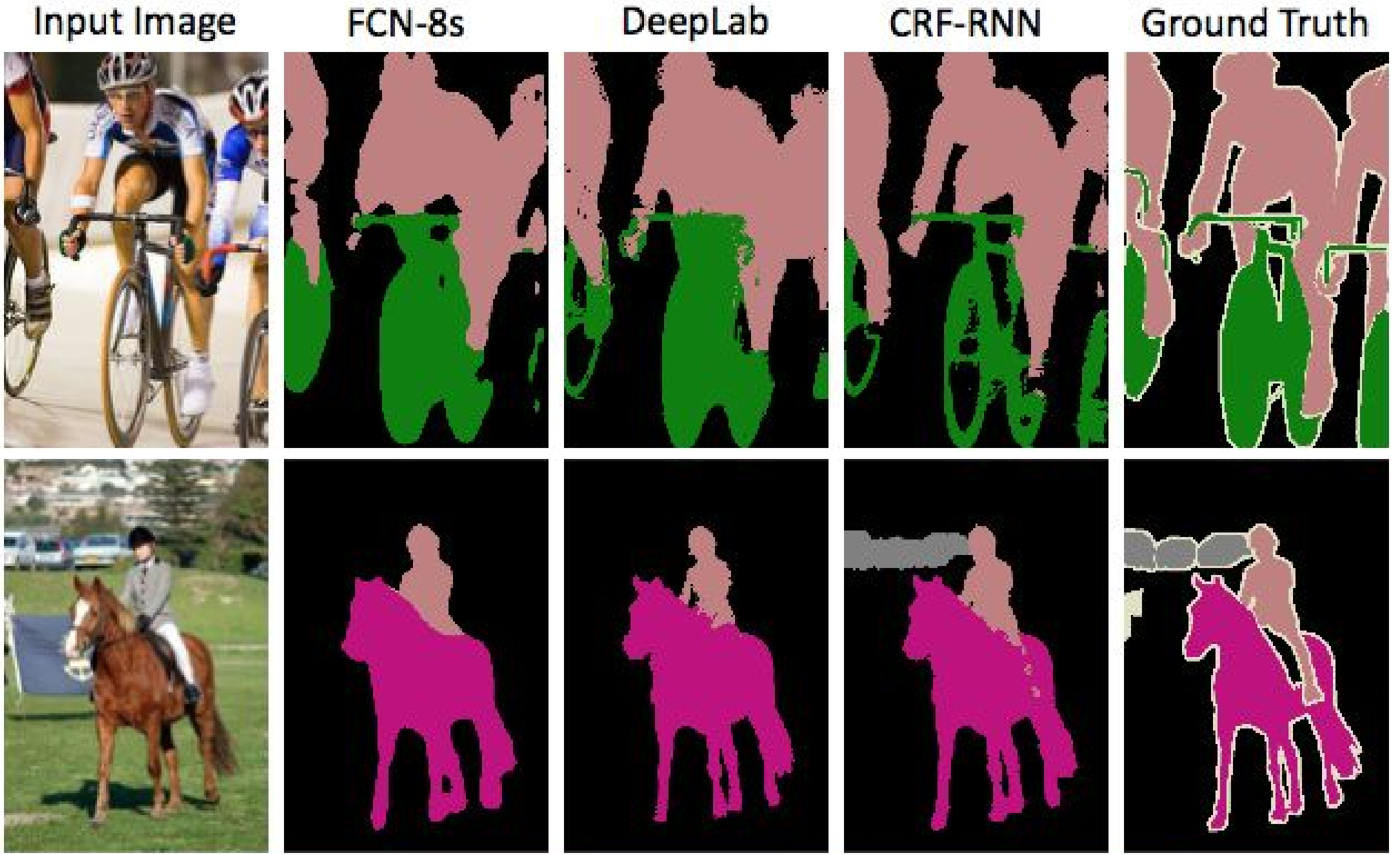
Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs.
Chen* & Papandreou* et al. ICLR 2015.

fully conv. nets + random fields



Method	Without COCO	With COCO
Plain FCN-8s	61.3	68.3
FCN-8s and CRF disconnected	63.7	69.5
End-to-end training of CRF-RNN	69.6	72.9

Conditional Random Fields as Recurrent Neural Networks. *Zheng* & Jayasumana* et al.* arxiv 2015.

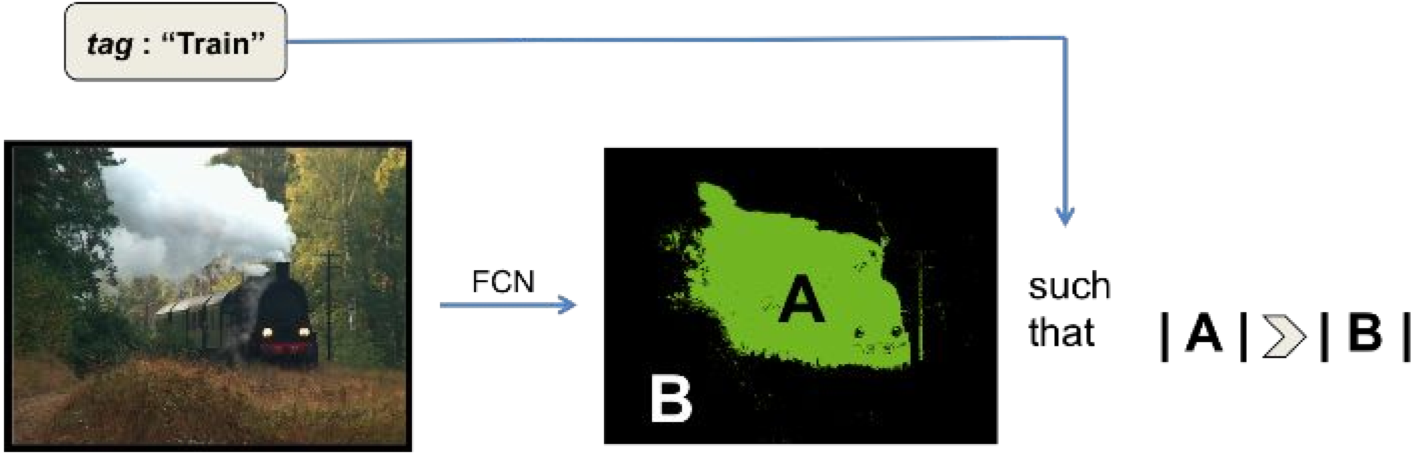


[comparison credit: CRF as RNN, Zheng* & Jayasumana* et al. ICCV 2015]

DeepLab: Chen* & Papandreou* et al. ICLR 2015. **CRF-RNN**: Zheng* & Jayasumana* et al. ICCV 2015

fully conv. nets + weak supervision

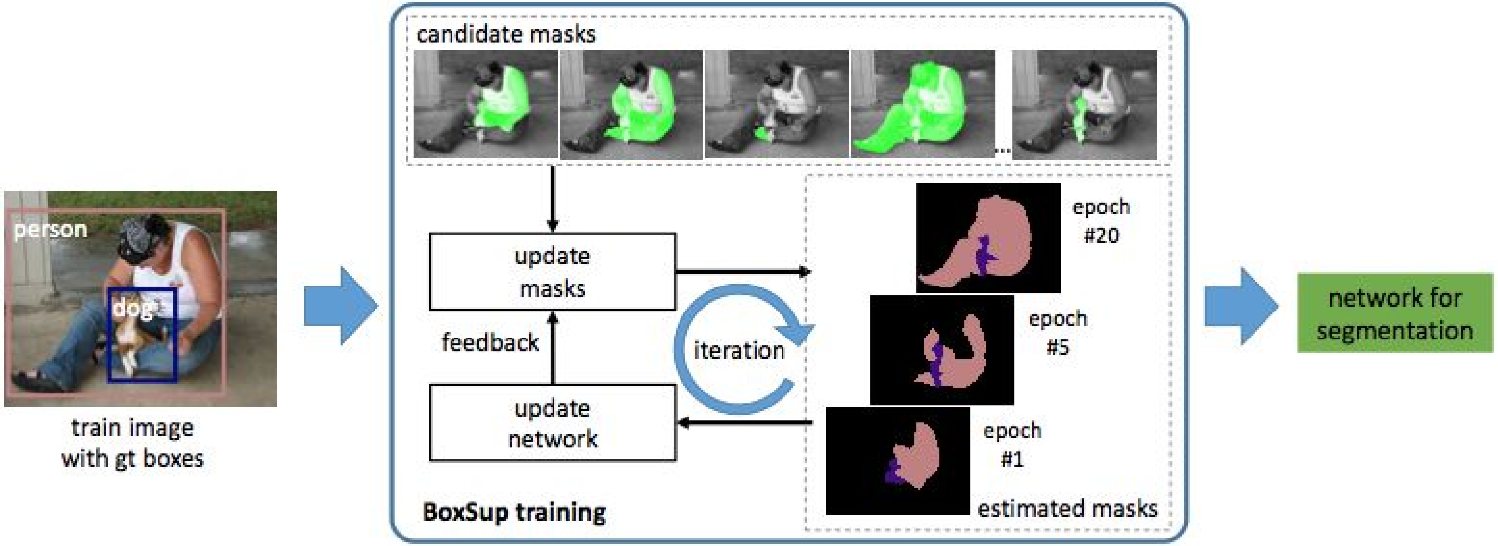
FCNs expose a spatial loss map to guide learning:
segment from tags by MIL or pixelwise constraints.



Constrained Convolutional Neural Networks for Weakly Supervised Segmentation.
Pathak et al. arXiv 2015.

fully conv. nets + weak supervision

FCNs expose a spatial loss map to guide learning:
mine boxes + feedback to refine masks.

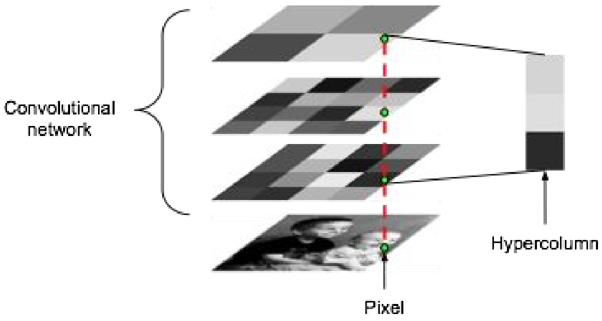


BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation.
Dai et al. 2015.

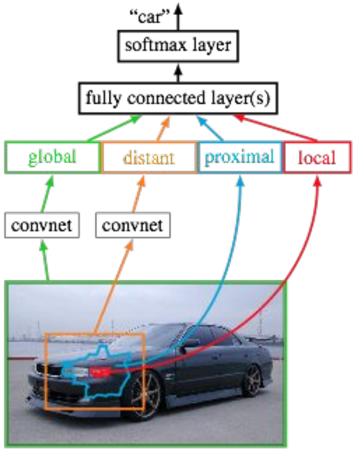
		mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
MSRA_BoxSup [?]	FCN	75.2	89.8	38.0	89.2	68.9	68.0	89.6	83.0	87.7	34.4	83.6	67.1	81.5	83.7	85.2	83.5	58.6	84.9	55.8	81.2	70.7	18-May-2015
Oxford_TVG_CRF_RNN_COCO [?]	FCN	74.7	90.4	55.3	88.7	68.4	69.8	88.3	82.4	85.1	32.6	78.5	64.4	79.6	81.9	86.4	81.8	58.6	82.4	53.5	77.4	70.1	22-Apr-2015
DeepLab-MSc-CRF-LargeFOV-COCO-CrossJoint [?]	FCN	73.9	89.2	46.7	88.5	63.5	68.4	87.0	81.2	86.3	32.6	80.7	62.4	81.0	81.3	84.3	82.1	56.2	84.6	58.3	76.2	67.2	26-Apr-2015
Adelaide_Context_CNN_CRF_VOC [?]	FCN	72.9	89.7	37.6	77.4	62.1	72.9	88.1	84.8	81.9	34.4	80.0	55.9	79.3	82.3	84.0	82.9	59.7	82.8	54.1	77.5	70.3	25-May-2015
DeepLab-CRF-COCO-LargeFOV [?]	FCN	72.7	89.1	38.3	88.1	63.3	69.7	87.1	83.1	85.0	29.3	76.5	56.5	79.8	77.9	85.8	82.4	57.4	84.3	54.9	80.5	64.1	18-Mar-2015
POSTECH_EDeconvNet_CRF_VOC [?]	FCN	72.5	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	22-Apr-2015
Oxford_TVG_CRF_RNN_VOC [?]	FCN	72.0	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	22-Apr-2015
DeepLab-MSc-CRF-LargeFOV [?]	FCN	71.6	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	02-Apr-2015
MSRA_BoxSup [?]	FCN	71.0	86.4	35.5	79.7	65.2	65.2	84.3	78.5	83.7	30.5	76.2	62.6	79.3	76.1	82.1	81.3	57.0	78.2	55.0	72.5	68.1	10-Feb-2015
DeepLab-CRF-COCO-Strong [?]	FCN	70.4	85.3	36.2	84.8	61.2	67.5	84.6	81.4	81.0	30.8	73.8	53.8	77.5	76.5	82.3	81.6	56.3	78.9	52.3	76.6	63.3	11-Feb-2015
DeepLab-CRF-LargeFOV [?]	FCN	70.3	83.5	36.6	82.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	48.8	[illegible]	[illegible]	28-Mar-2015
TTI_zoomout_v2 [?]		69.6	85.6	37.3	83.2	62.5	66.0	85.1	80.7	84.9	27.2	73.2	57.5	78.1	79.2	81.1	77.1	53.6	74.0	49.2	71.7	63.3	30-Mar-2015
DeepLab-CRF-MSc [?]	FCN	67.1	80.4	36.8	77.4	55.2	66.4	81.5	77.5	78.9	27.1	68.2	52.7	74.3	69.6	79.4	79.0	56.9	78.8	45.2	72.7	59.3	30-Dec-2014
DeepLab-CRF [?]	FCN	66.4	78.4	33.1	78.2	55.6	65.3	81.3	75.5	78.6	25.3	69.2	52.7	75.2	69.0	79.1	77.6	54.7	78.3	45.1	73.3	56.2	23-Dec-2014
CRF_RNN [?]	FCN	65.2	80.9	34.0	72.9	52.6	62.5	79.8	76.3	79.9	23.6	67.7	51.8	74.8	69.9	76.9	76.9	49.0	74.7	42.7	72.1	59.6	10-Feb-2015
TTI_zoomout_16 [?]		64.4	81.9	35.1	78.2	57.4	56.5	80.5	74.0	79.8	22.4	69.6	53.7	74.0	76.0	76.6	68.8	44.3	70.2	40.2	68.9	55.3	24-Nov-2014
Hypercolumn [?]		62.6	68.7	33.5	69.8	51.3	70.2	81.1	71.9	74.9	23.9	60.6	46.9	72.1	68.3	74.5	72.9	52.6	64.4	45.4	64.9	57.4	09-Apr-2015
FCN-8s [?]	FCN	62.2	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	12-Nov-2014
MSRA_CFM [?]		61.8	75.7	26.7	69.5	48.8	65.6	81.0	69.2	73.3	30.0	68.7	51.5	69.1	68.1	71.7	67.5	50.4	66.5	44.4	58.9	53.5	17-Dec-2014
TTI_zoomout [?]		58.4	70.3	31.9	68.3	46.4	52.1	75.3	68.4	75.3	19.2	58.4	49.9	69.6	63.0	70.1	67.6	41.5	64.0	34.9	64.2	47.3	17-Nov-2014
SDS [?]		51.6	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	21-Jul-2014
NUS_UDS [?]		50.0	67.0	24.5	47.2	45.0	47.9	65.3	60.6	58.5	15.5	50.8	37.4	45.8	59.9	62.0	52.7	40.8	48.2	36.8	53.1	45.6	29-Oct-2014
TTIC-divmbest-rerank [?]		48.1	62.7	25.6	46.9	43.0	54.8	58.4	58.6	55.6	14.6	47.5	31.2	44.7	51.0	60.9	53.5	36.6	50.9	30.1	50.2	46.8	15-Nov-2012
BONN_O2PCPMC_FGT_SEGM [?]		47.8	64.0	27.3	54.1	39.2	48.7	56.6	57.7	52.5	14.2	54.8	29.6	42.2	58.0	54.8	50.2	36.6	58.6	31.6	48.4	38.6	08-Aug-2013
BONN_O2PCPMC_FGT_SEGM [?]		47.5	63.4	27.3	56.1	37.7	47.2	57.9	59.3	55.0	11.5	50.8	30.5	45.0	58.4	57.4	48.6	34.6	53.3	32.4	47.6	39.2	23-Sep-2012
BONNGC_O2P_CPMC_CSI [?]		46.8	63.6	26.8	45.6	41.7	47.1	54.3	58.6	55.1	14.5	49.0	30.9	46.1	52.6	58.2	53.4	32.0	44.5	34.6	45.3	43.1	23-Sep-2012
BONN_CMBR_O2P_CPMC_LIN [?]		46.7	63.9	23.8	44.6	40.3	45.5	59.6	58.7	57.1	11.7	45.9	34.9	43.0	54.9	58.0	51.5	34.6	44.1	29.9	50.5	44.5	23-Sep-2012

== segmentation with Caffe

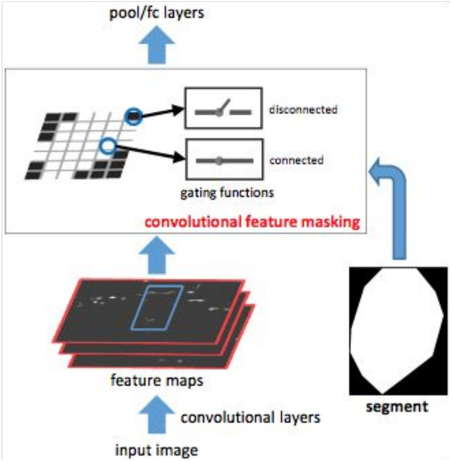
caffeinated contemporaries



Hypercolumn SDS
Hariharan, Arbeláez, Girshick, Malik



Zoom-Out
Mostajabi, Yadollahpour, Shaknarovich



Convolutional Feature Masking
Dai, He, Sun

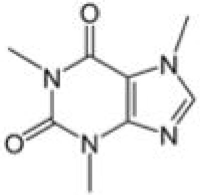
conclusion

fully convolutional networks are fast, end-to-end models for pixelwise problems

- **code** in Caffe branch (merged soon)
- **models** for PASCAL VOC, NYUDv2, SIFT Flow, PASCAL-Context

fcn.berkeleyvision.org

model example
inference example
solving example



caffe.berkeleyvision.org

github.com/BVLC/caffe