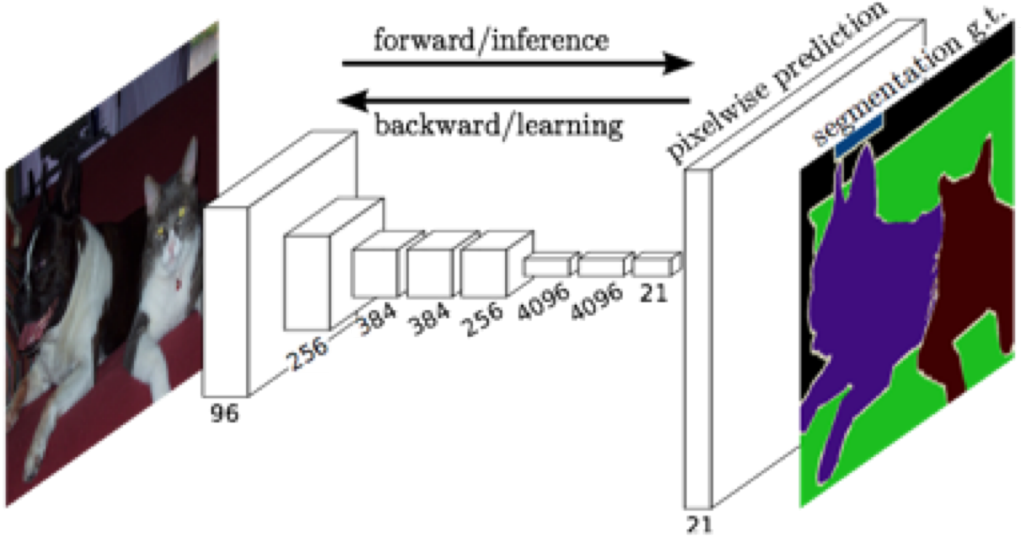
Fully Convolutional Networks for Semantic Segmentation



Jonathan Long* Evan Shelhamer* Trevor Darrell

UC Berkeley

Presented by: Akhilesh Kumar

Slide credit: Jonathan Long

Overview

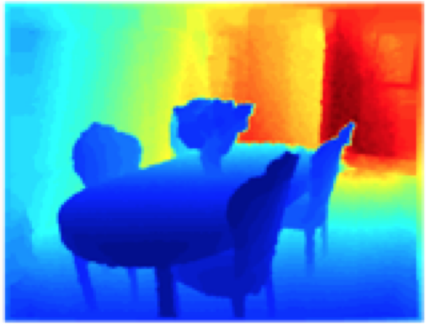
- Reinterpret standard classification convnets as “Fully convolutional” networks (FCN) for semantic segmentation
- Use AlexNet, VGG, and GoogleNet in experiments
- Novel architecture: combine information from different layers for segmentation
- State-of-the-art segmentation for PASCAL VOC 2011/2012, NYUDv2, and SIFT Flow at the time
- Inference less than one fifth of a second for a typical image

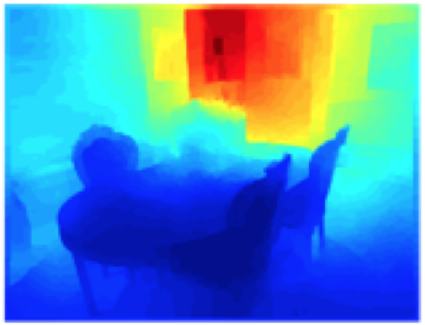
Slide credit: Jonathan Long

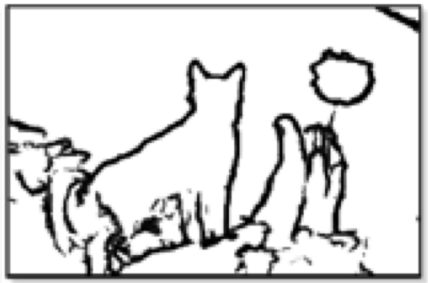
pixels in, pixels out

semantic segmentation

monocular depth estimation (Liu et al. 2015)



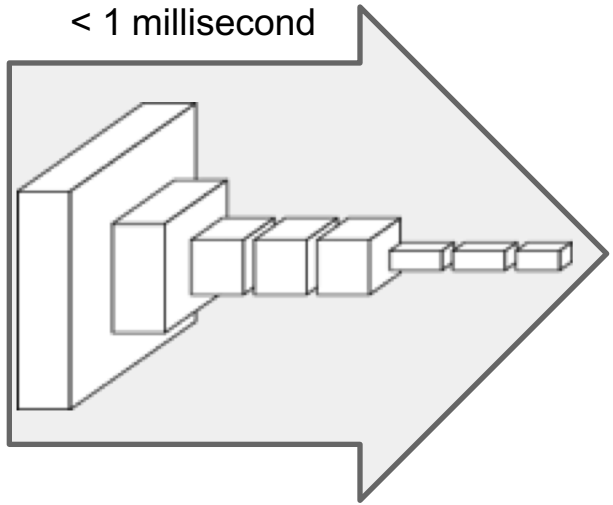


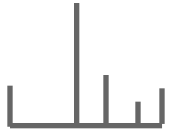


boundary prediction (Xie & Tu 2015)

Slide credit: Jonathan Long

convnets perform classification





“tabby cat”

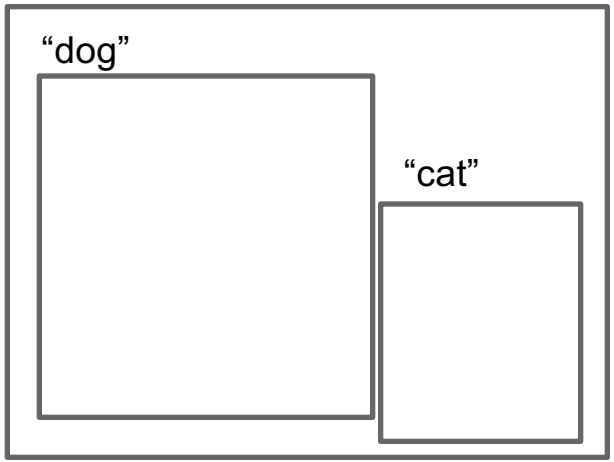
1000-dim vector

Slide credit: Jonathan Long

R-CNN does detection

many seconds

R-CNN



Slide credit: Jonathan Long

R-CNN

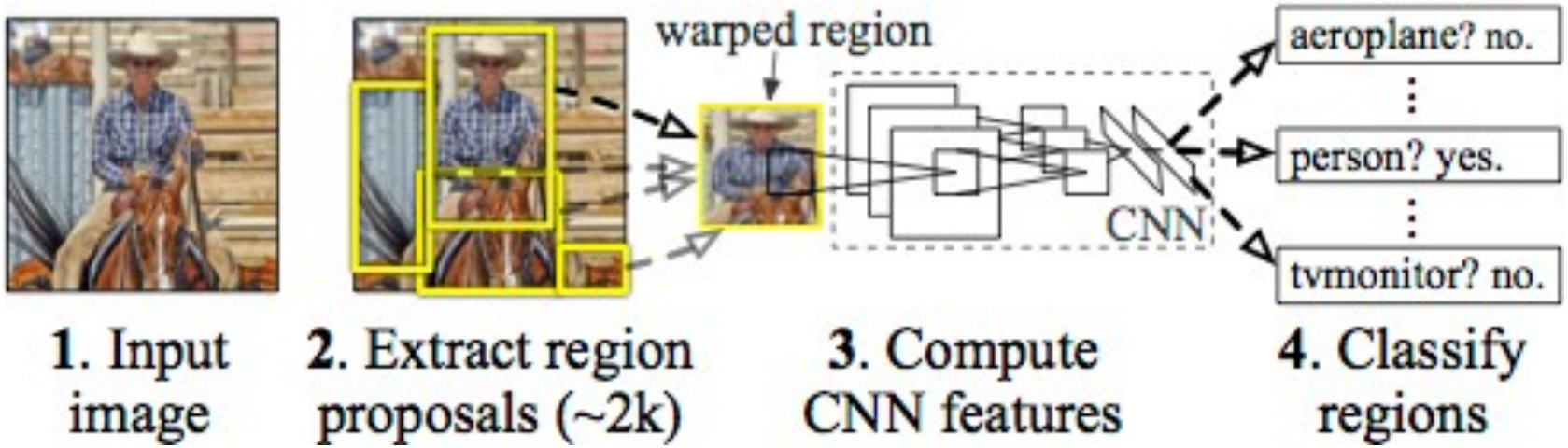


figure: Girshick et al.

Slide credit: Jonathan Long

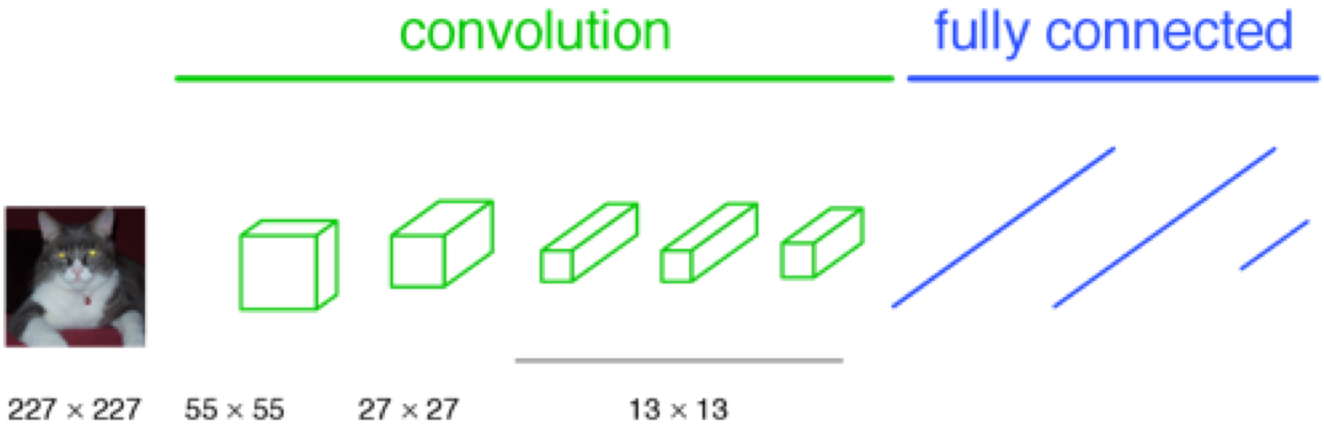
< 1/5 second

???

end-to-end learning

Slide credit: Jonathan Long

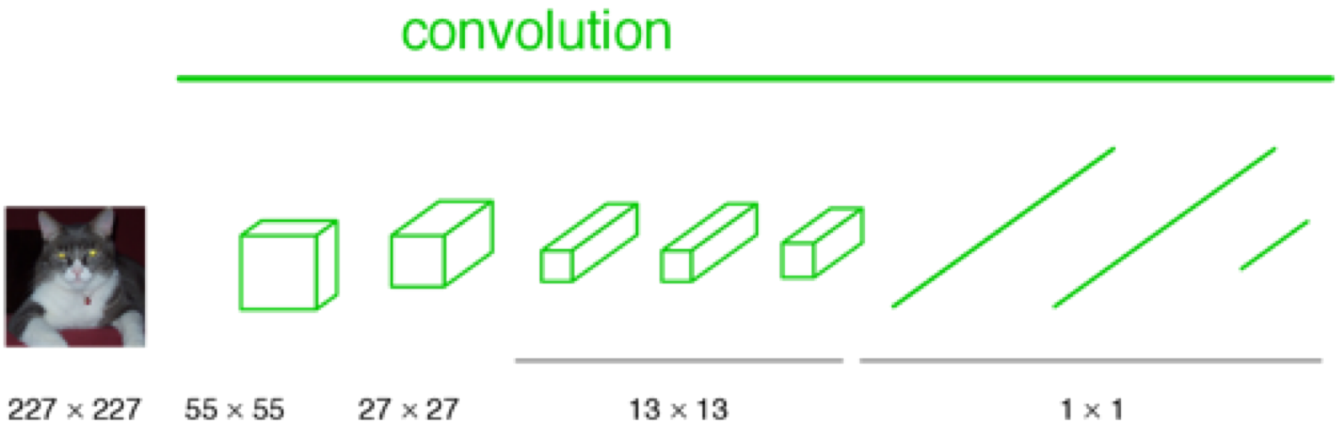
a classification network



“tabby cat”

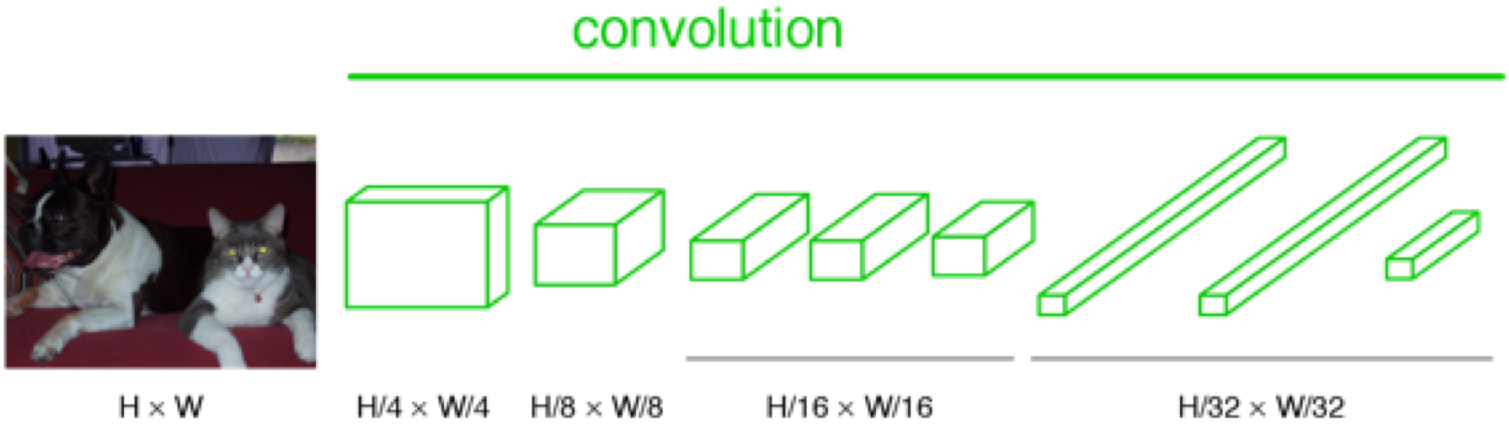
Slide credit: Jonathan Long

becoming fully convolutional



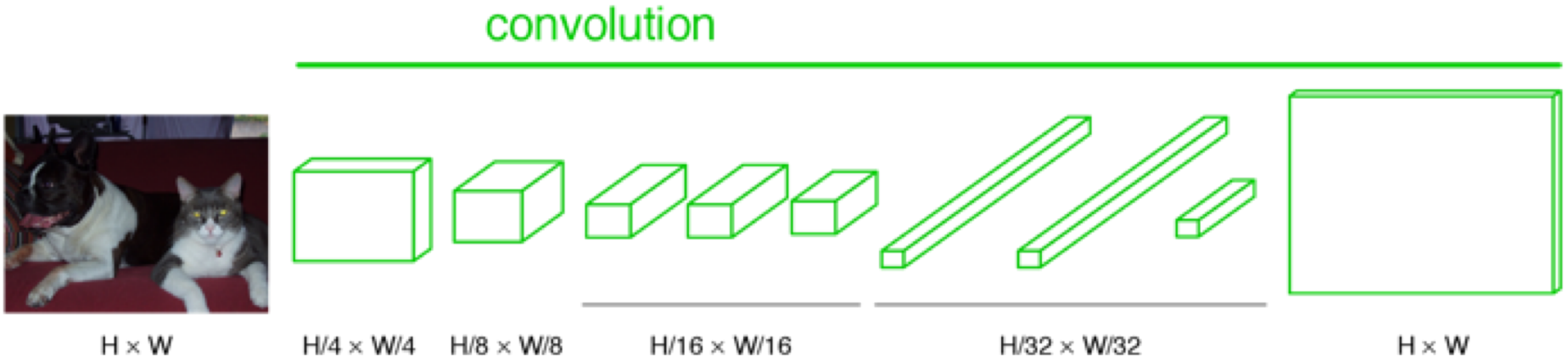
Slide credit: Jonathan Long

becoming fully convolutional



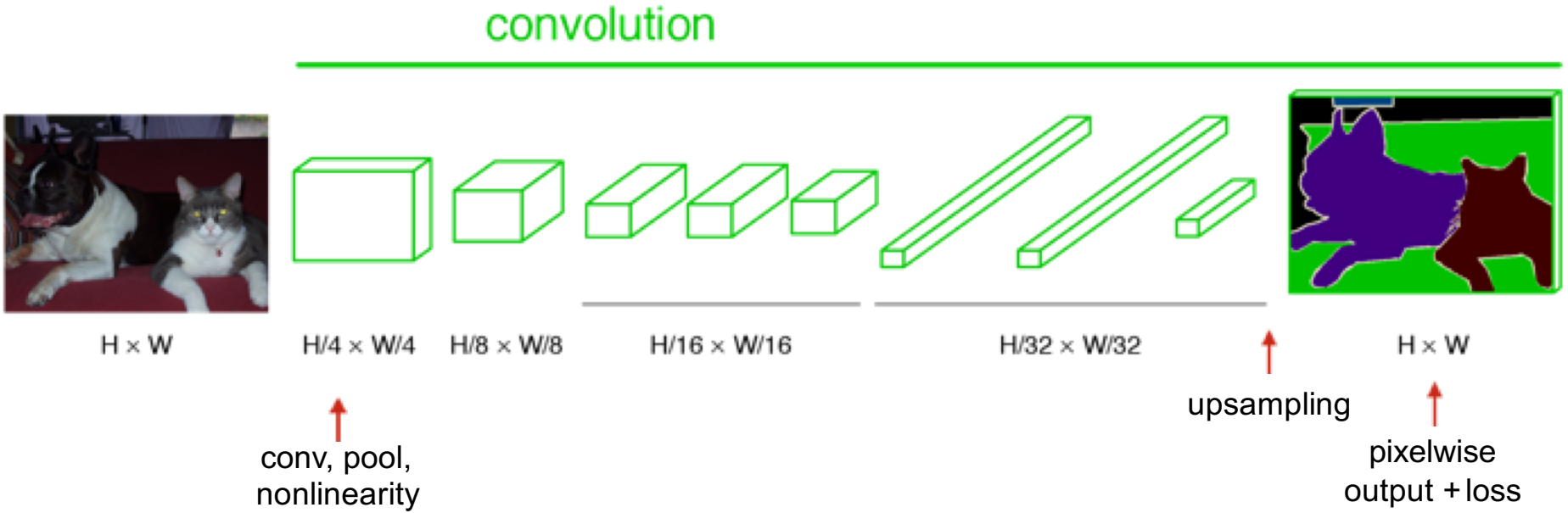
Slide credit: Jonathan Long

upsampling output



Slide credit: Jonathan Long

network



Slide credit: Jonathan Long

Dense Predictions

- Shift-and-stitch: trick that yields dense predictions without interpolation
- Upsampling via deconvolution
- Shift-and-stitch used in preliminary experiments, but not included in final model
- Upsampling found to be more effective and efficient

Classifier to Dense FCN

- Convolutionalize proven classification architectures: AlexNet, VGG, and GoogLeNet (reimplementation)
- Remove classification layer and convert all fully connected layers to convolutions
- Append 1x1 convolution with channel dimensions and predict scores at each of the coarse output locations (21 categories + background for PASCAL)

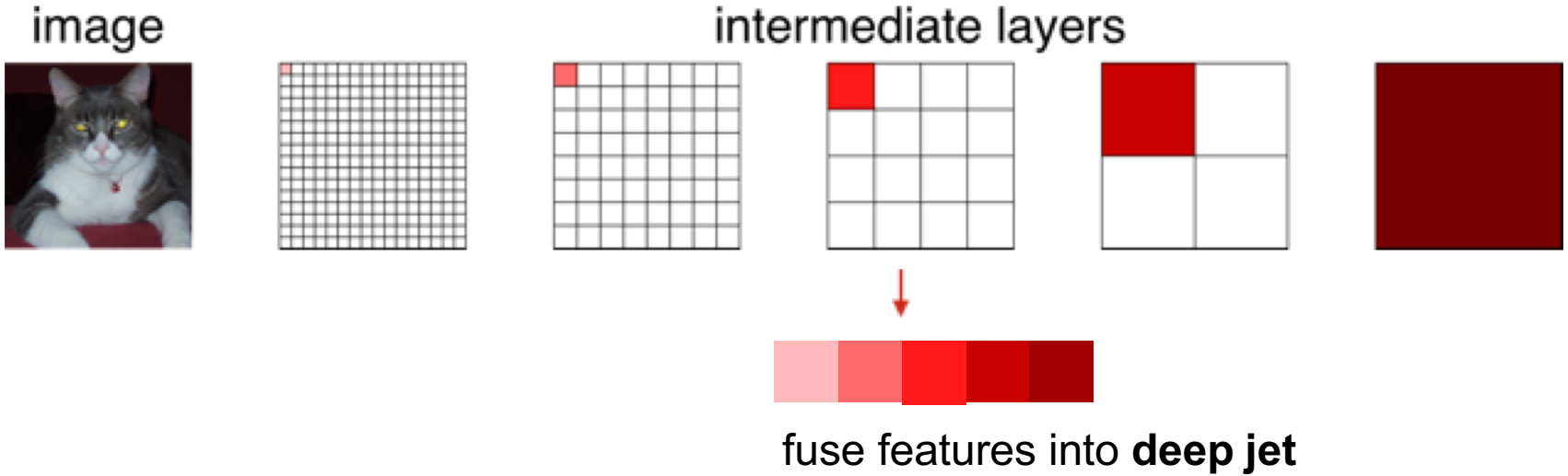
Classifier to Dense FCN

Cast ILSVRC classifiers into FCNs and compare performance on validation set of PASCAL 2011

	FCN-AlexNet	FCN-VGG16	FCN-GoogLeNet[4]
mean IU	39.8	**56.0**	42.5
forward time	50 ms	210 ms	59 ms
conv. layers	8	16	22
parameters	57M	134M	6M
rf size	355	404	907
max stride	32	32	32

spectrum of deep features

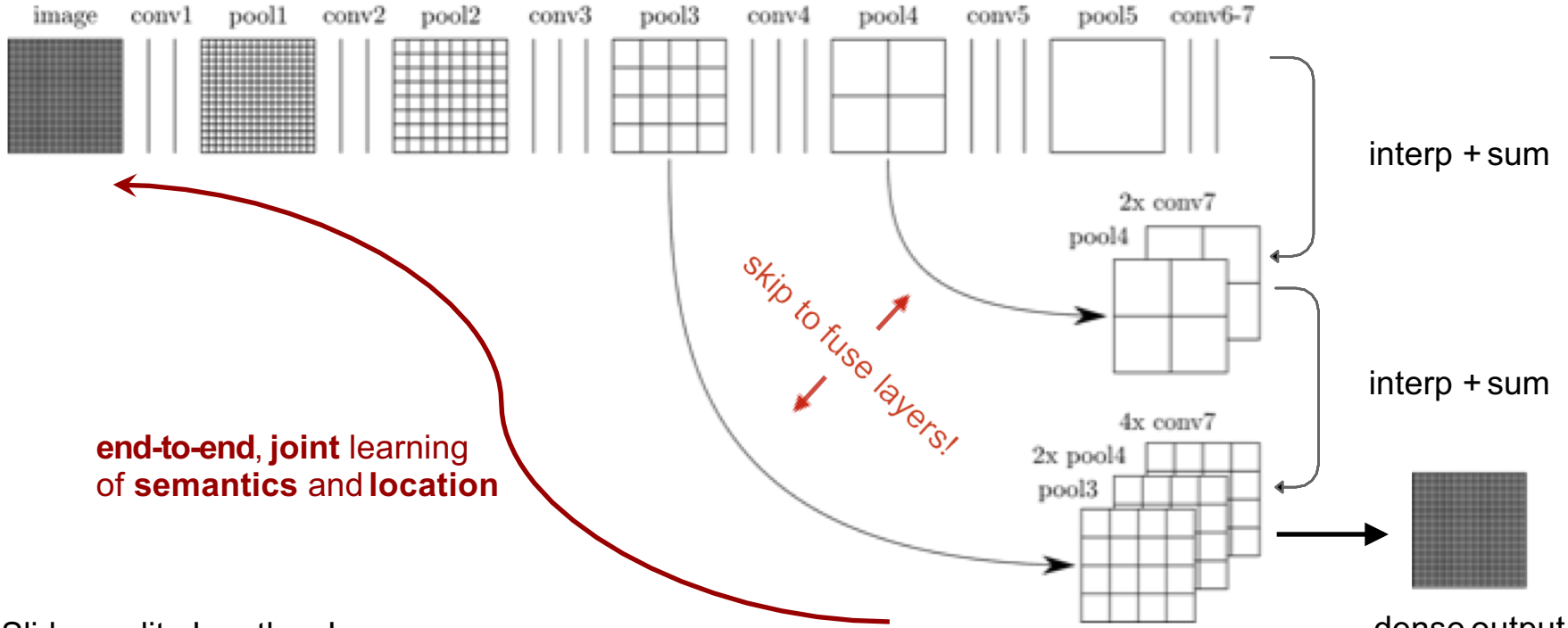
combine *where* (local, shallow) with *what* (global, deep)



Slide credit: Jonathan Long

(cf. Hariharan et al. CVPR15 "hypercolumn")

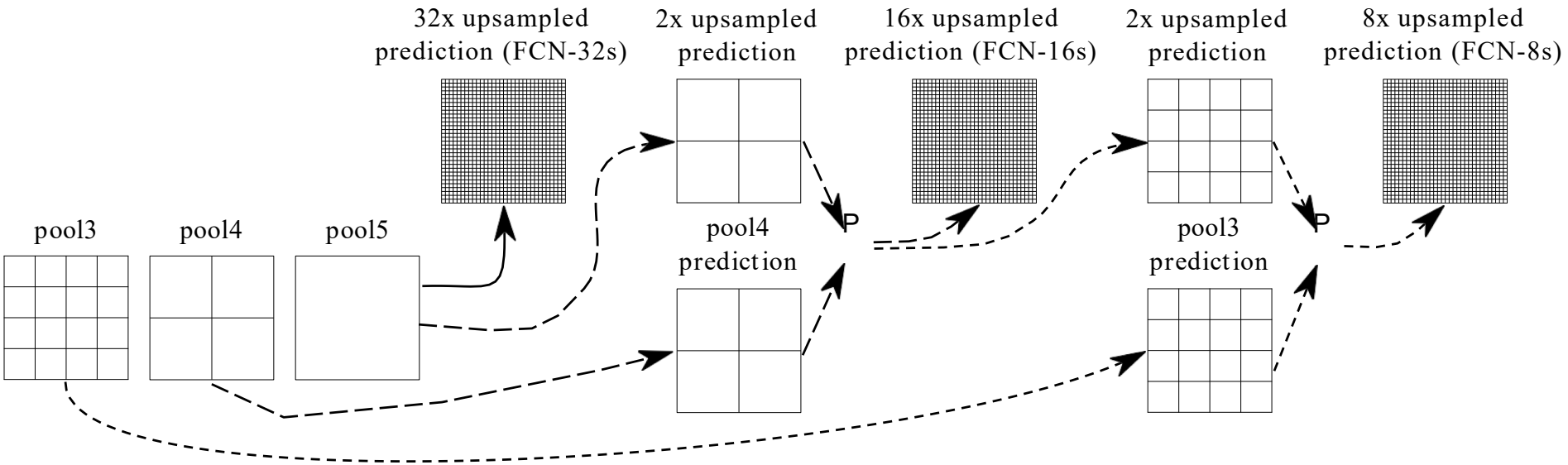
skip layers



end-to-end, **joint** learning
of **semantics** and **location**

Slide credit: Jonathan Long

skip layers

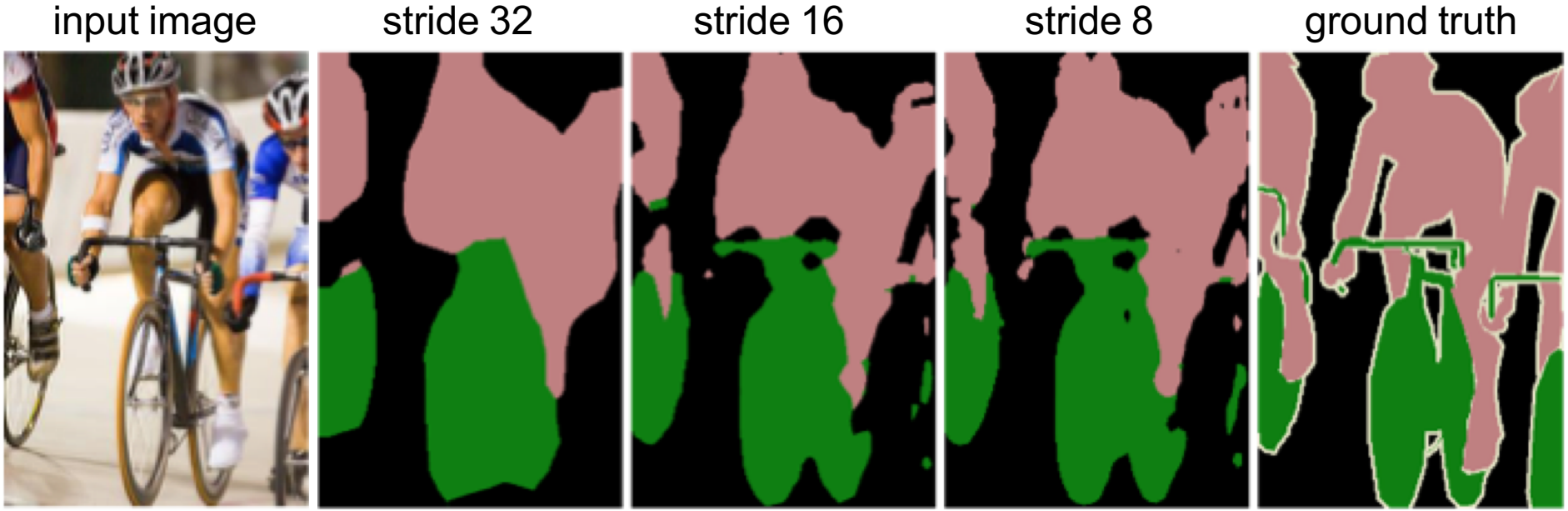

32x upsampled prediction (FCN-32s)
2x upsampled prediction
16x upsampled prediction (FCN-16s)
2x upsampled prediction
8x upsampled prediction (FCN-8s)
pool3
pool4
pool5
pool4 prediction
pool3 prediction
P
P

Comparison of skip FCNs

Results on subset of validation set of PASCAL VOC 2011

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**

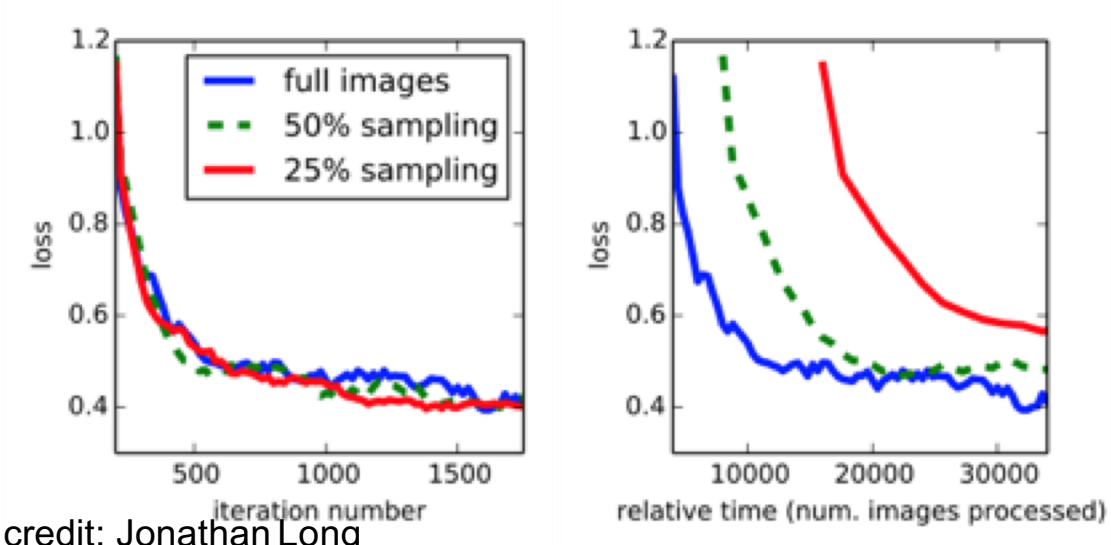
skip layer refinement



Slide credit: Jonathan Long

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



Slide credit: Jonathan Long

Results – PASCAL VOC 2011/12

VOC 2011: 8498 training images (from additional labeled data

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [16]	52.6	51.6	⇠ 50 s
FCN-8s	**62.7**	**62.2**	⇠ **175 ms**

Results – NYUDv2

1449 RGB-D images with pixelwise labels →→40 categories

	pixel acc.	mean acc.	mean IU	f.w. IU
Gupta *et al.* [14]	60.3	-	28.6	47.0
FCN-32s RGB	60.0	42.2	29.2	43.9
FCN-32s RGBD	61.5	42.4	30.5	45.5
FCN-32s HHA	57.1	35.2	24.2	40.4
FCN-32s RGB-HHA	64.3	44.9	32.8	48.0
FCN-16s RGB-HHA	**65.4**	**46.1**	**34.0**	**49.5**

Results – SIFT Flow

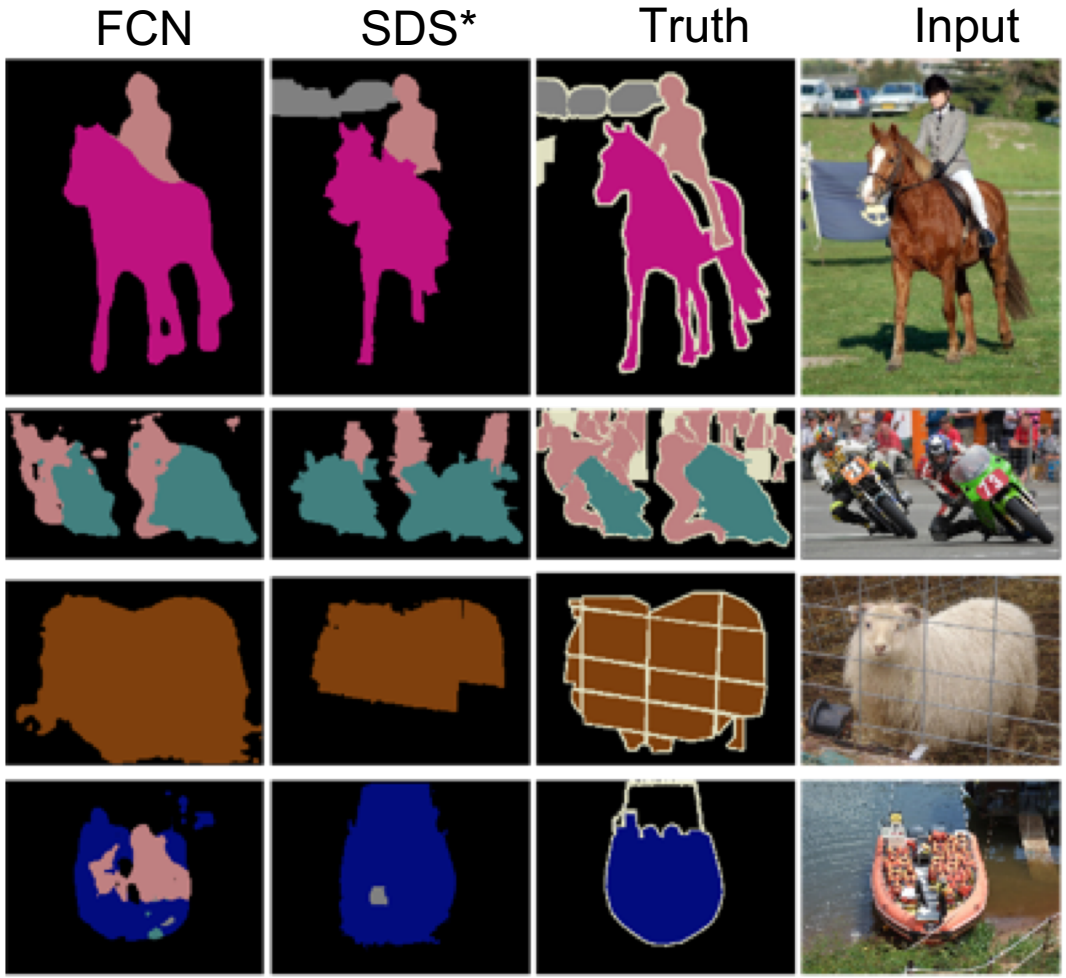
2688 images with pixel labels

→→33 semantic categories, 3 geometric categories

Learn both label spaces jointly

→→learning and inference have similar performance and computation as independent models

	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [23]	76.7	-	-	-	-
Tighe *et al.* [33]	-	-	-	-	90.8
Tighe *et al.* [34]1	75.6	41.1	-	-	-
Tighe *et al.* [34]2	78.6	39.2	-	-	-
Farabet *et al.* [8]1	72.3	50.8	-	-	-
Farabet *et al.* [8]2	78.5	29.6	-	-	-
Pinheiro *et al.* [28]	77.7	29.8	-	-	-
FCN-16s	**85.2**	**51.7**	39.5	76.1	**94.3**



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

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

Slide credit: Jonathan Long

*Simultaneous Detection and Segmentation Hariharan et al. ECCV14

== segmentation with Caffe

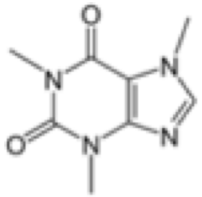
		mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
MSRA_BoxSup [?]	FCN	75.2	89.8	38.0	89.2	68.9	68.0	89.6	83.0	87.7	34.4	83.6	67.1	81.5	83.7	85.2	83.5	58.6	84.9	55.8	81.2	70.7	18-May-2015
Oxford_TVG_CRF_RNN_COCO [?]	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-CrossJo[illegible]	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	[illegible]	82.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	28-Mar-2015
TTI_zoomout_v2 [?]		69.6	85.6	37.3	83.2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	27.2	[illegible]	[illegible]	[illegible]	[illegible]	81.1	77.1	53.6	[illegible]	49.2	[illegible]	[illegible]	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
[illegible]		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	[illegible]-2012

Slide credit: Jonathan Long

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



caffe.berkeleyvision.org

github.com/BVLC/caffe

fcn.berkeleyvision.org

Slide credit: Jonathan Long