Fully Convolutional Networks for Semantic Segmentation

Evan Shelhamer , Jonathan Long , and Trevor Darrel
UC Berkeley

Presented by: Martin Cote

Prepared for: ME780 Perception for Autonomous Driving

UNIVERSITY OF WATERLOO

Overview

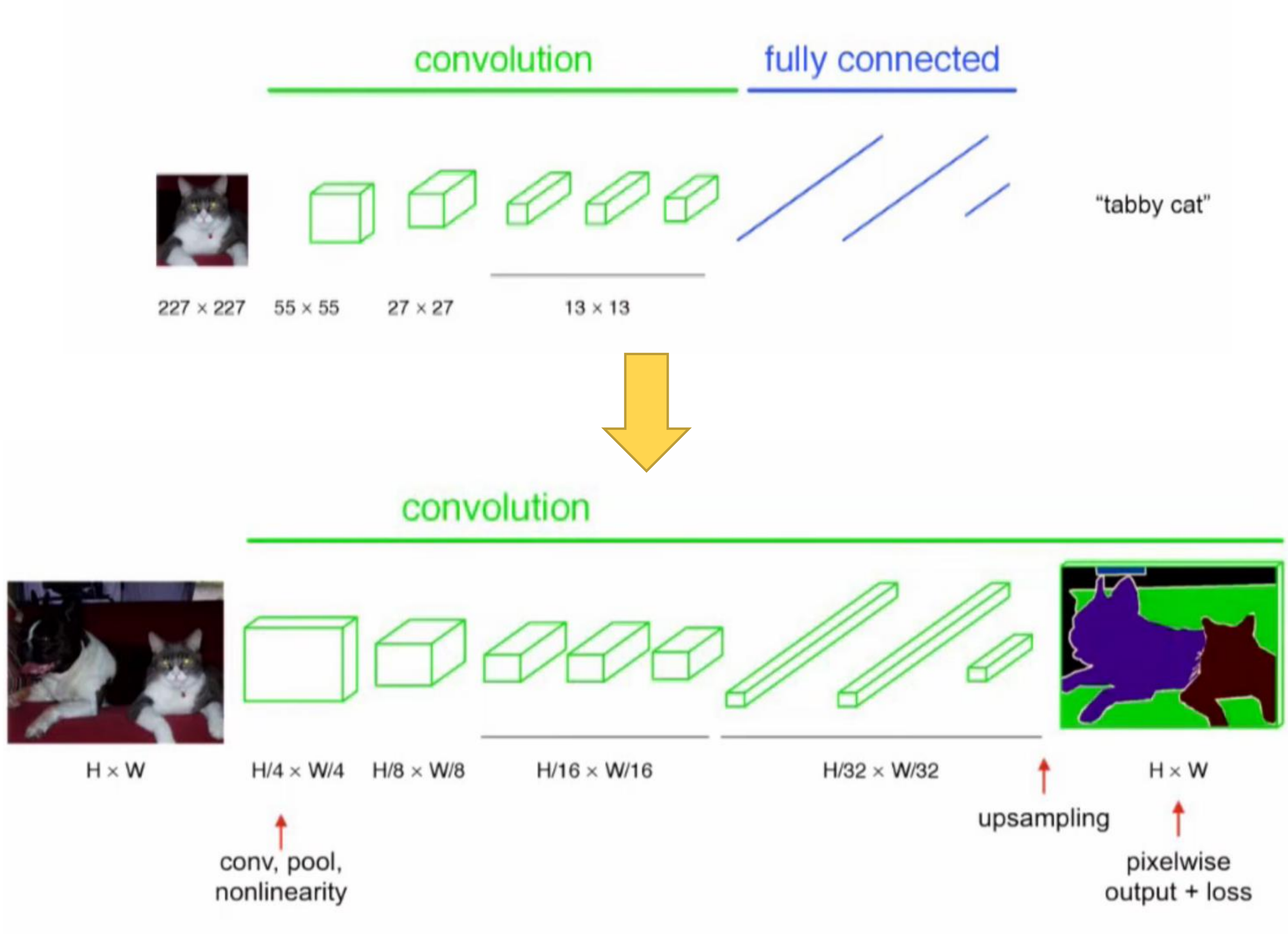
- Motivation
- Network Architecture
  - Fully convolutional networks
  - Skip layers
- Results
- Summary

Motivation

- Use convnets to make pixel-wise predictions
- Semantic segmentation provides the “what” and “where”
- Does not require
  - Patchwise training
  - Refinement by superpixel projection, random field regularization, filtering, or local classification
  - interlacing to obtain dense output
  - multi-scale pyramid
  - saturating tanh nonlinearities
  - ensembles

Network Architecture

Changes from a standard Convnet classifier

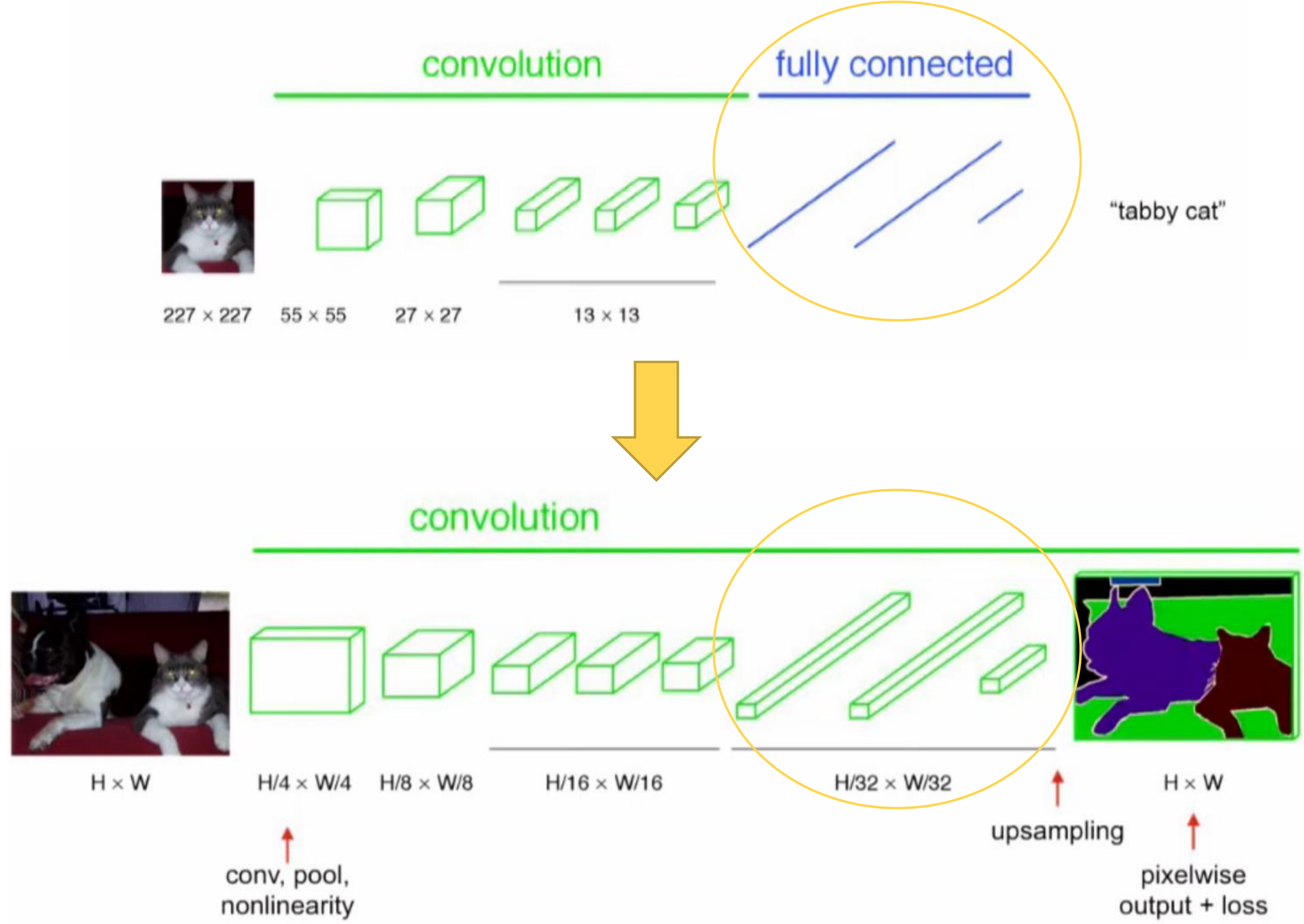


UNIVERSITY OF WATERLOO

Network Architecture

Changes from a standard Convnet classifier

1. Change fully connected layers to convolution layers with 1 x 1 output, so network is “fully convolutional, and no layers have pre-defined input size

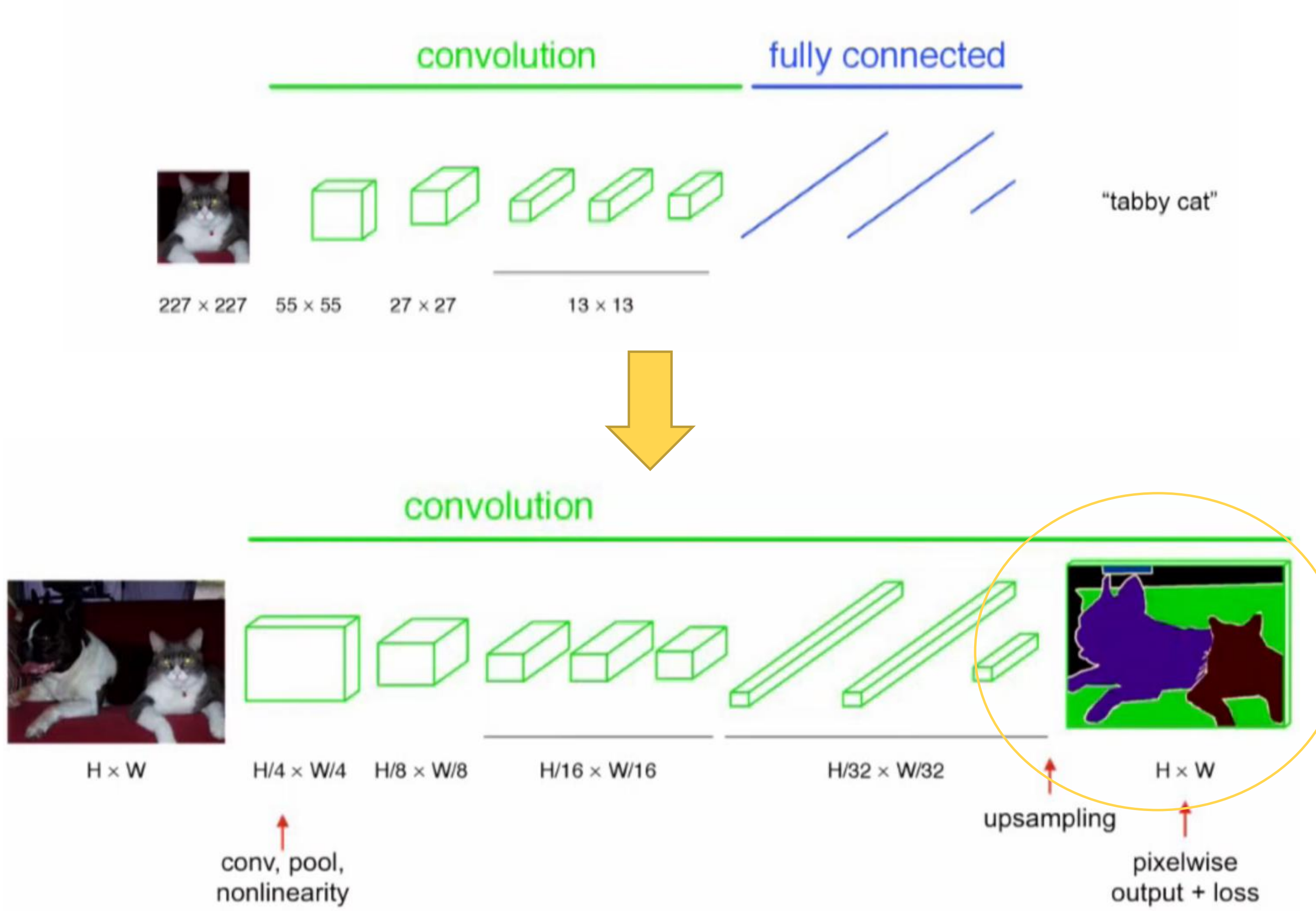


UNIVERSITY OF WATERLOO

Network Architecture

Changes from a standard Convnet classifier

1. Change fully connected layers to convolution layers with 1 x 1 output, so network is "fully convolutional, and no layers have pre-defined input size
2. Add an up sampling convolution layer, to get back to an output of the input's image H x W

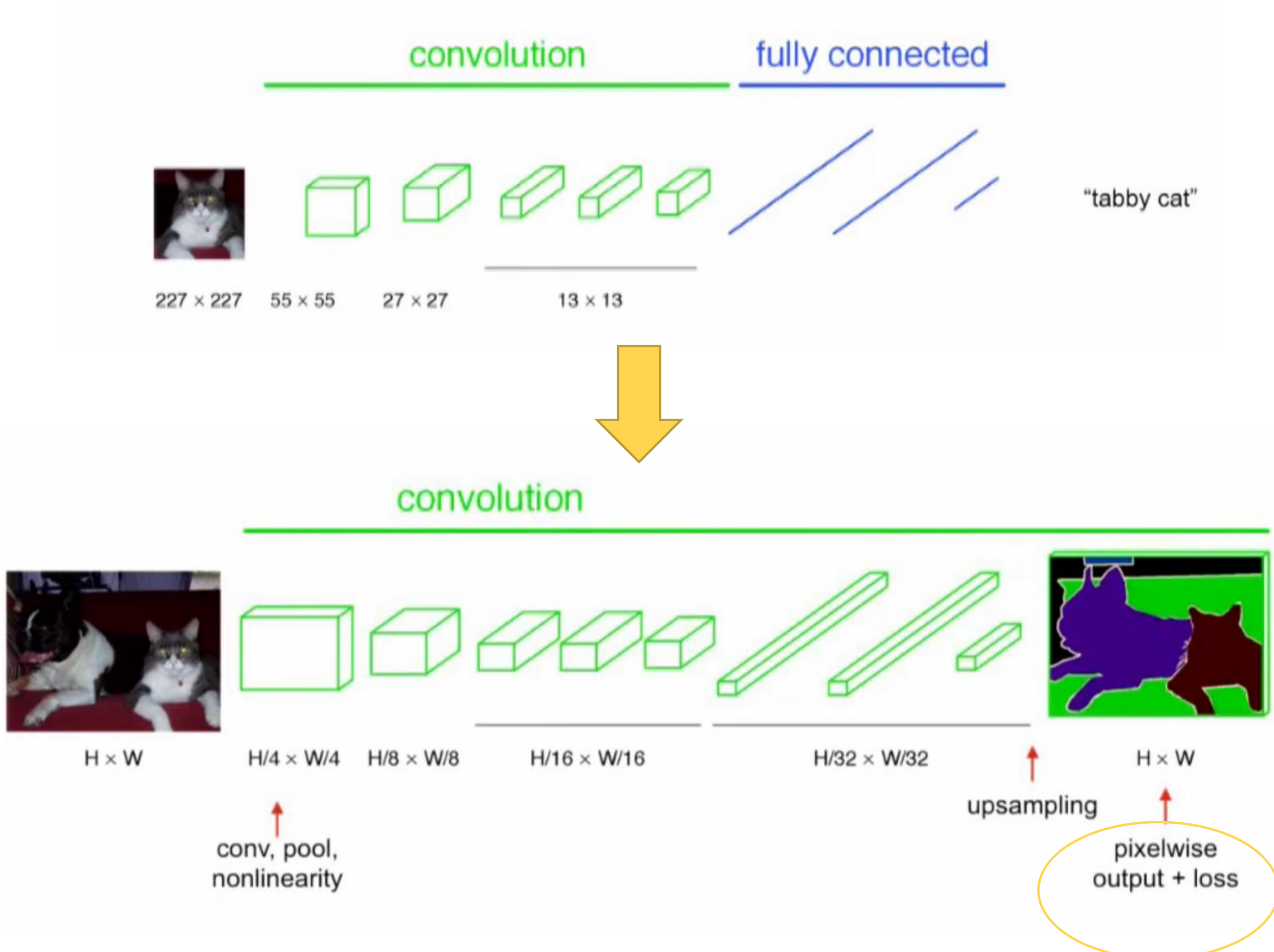


UNIVERSITY OF WATERLOO

Network Architecture

Changes from a standard Convnet classifier

1. Change fully connected layers to convolution layers with 1 x 1 output, so network is “fully convolutional, and no layers have pre-defined input size
2. Add an up sampling convolution layer, to get back to an output of the input’s image H x W
3. Pper-pixel softmax loss for end to end learning



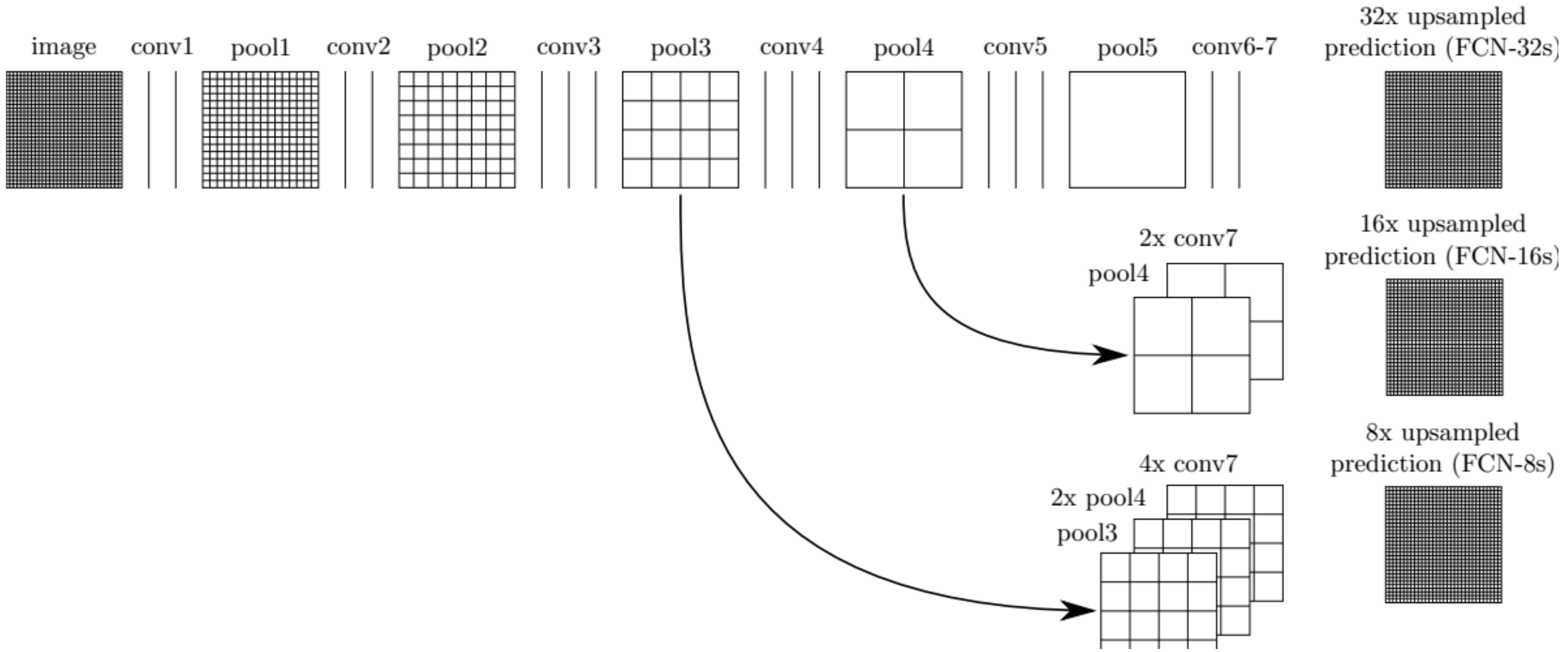
Network Architecture

This strategy was adopted to other well known networks, and tested on the PASCAL VOC2011 validation set

	FCN-AlexNet	FCN-VGG16	FCN-GoogLeNet[3]
mean IU	39.8	**56.0**	42.5
forward time	16 ms	100 ms	20 ms
conv. layers	8	16	22
parameters	57M	134M	6M
rf size	355	404	907
max stride	32	32	32

Skip Layers

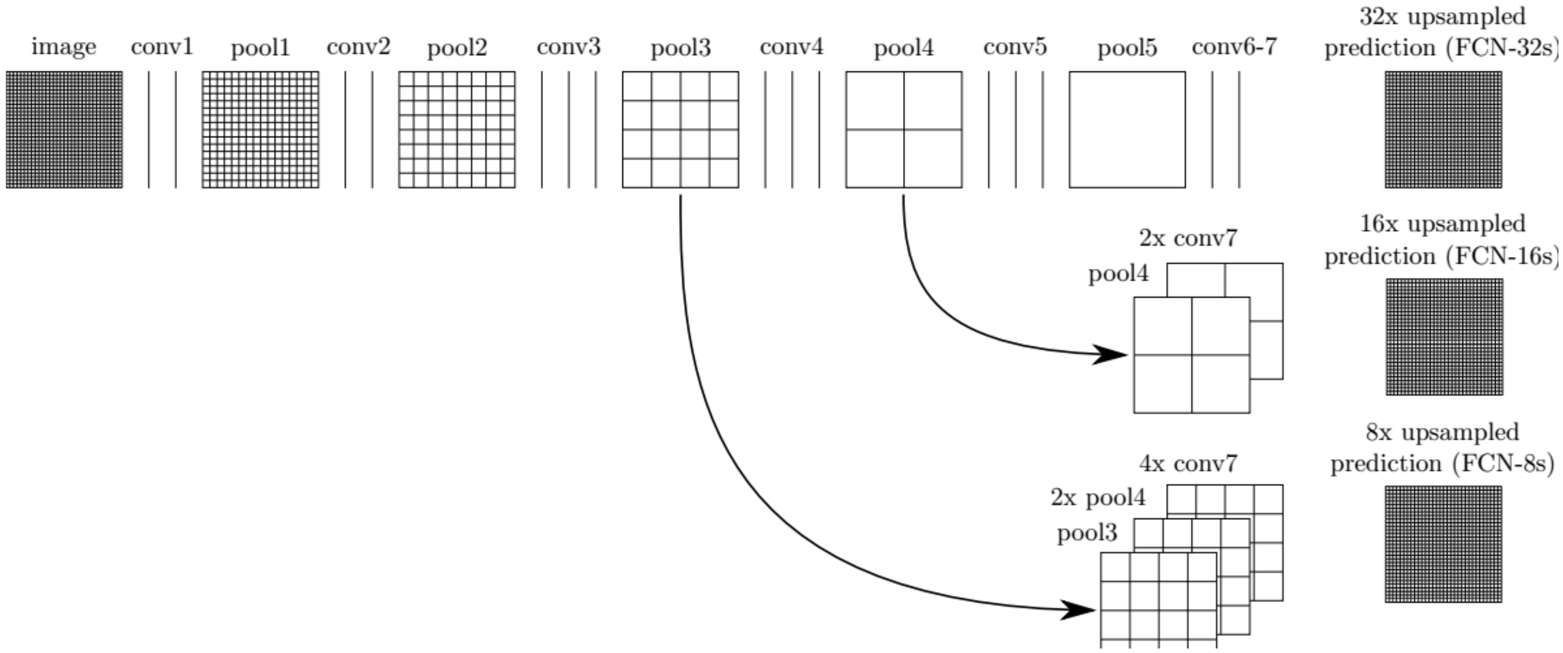
- Skip layers were introduced to combine
  - Shallow local layers which contains “where”
  - Deep global layers which contains “what”



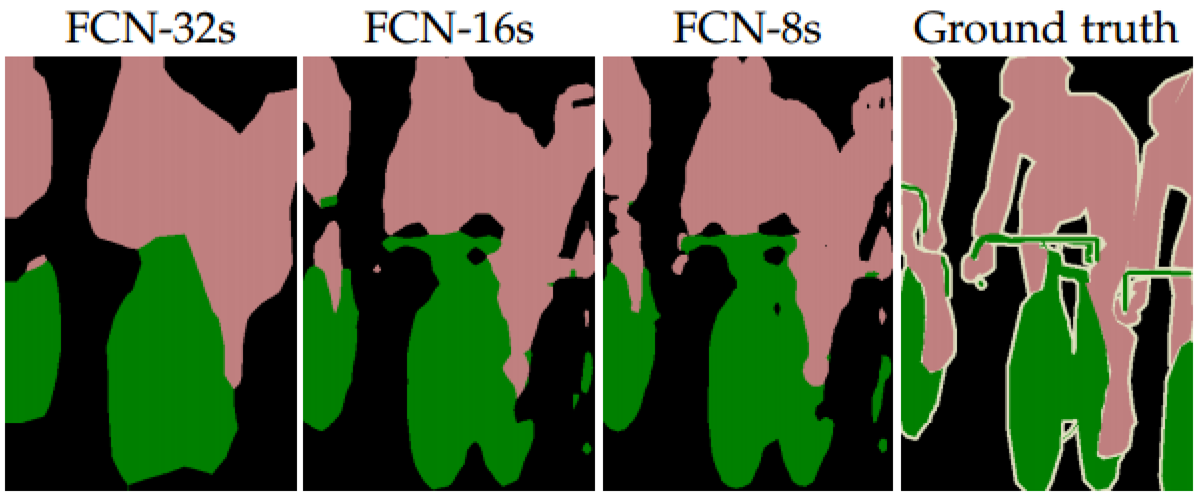
UNIVERSITY OF WATERLOO

Deep Jet

- Final layers to be fused are
  - aligned by scaling and cropping
  - Concatenated
  - Passed into 1 x 1 scoring layer



Results



FCN-32s = Fully convolutional version of VGG16
FCN-16s = Fully convolutional version of VGG16 with 1 skip layer
FCN-8s = Fully convolutional version of VGG16 with 2 skip layer

Results

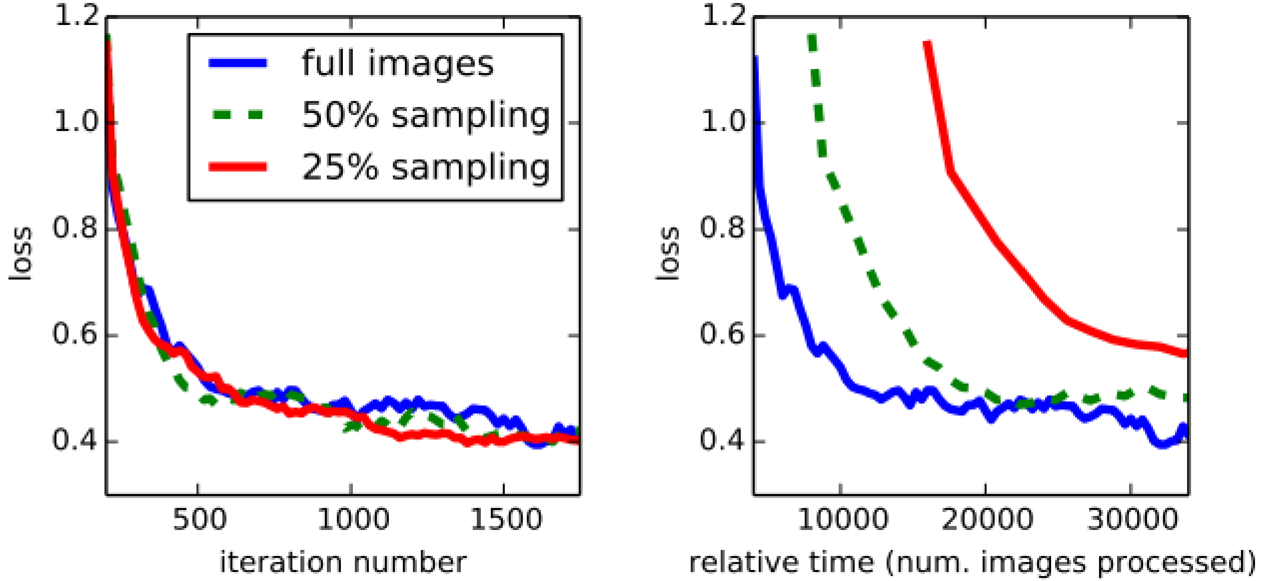
- Training the network in stages (adding 1 skip stream at a time) did not provide significant improvements over training all at once
- The paper conclude they've reached diminishing returns between FCN-16s and FCN-8s

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s	90.5	76.5	63.6	83.5
FCN-16s	91.0	78.1	65.0	84.3
FCN-8s at-once	91.1	**78.5**	65.4	84.4
FCN-8s staged	**91.2**	77.6	**65.5**	**84.5**
FCN-32s fixed	82.9	64.6	46.6	72.3
FCN-pool5	87.4	60.5	50.0	78.5
FCN-pool4	78.7	31.7	22.4	67.0
FCN-pool3	70.9	13.7	9.2	57.6

FCN-32s = Fully convolutional version of VGG16
FCN-16s = Fully convolutional version of VGG16 with 1 skip layer
FCN-8s = Fully convolutional version of VGG16 with 2 skip layer

Results

- Patch sampling is compared to full image training, and full image training converges quicker, with similar accuracy



FCN-32s = Fully convolutional version of VGG16
FCN-16s = Fully convolutional version of VGG16 with 1 skip layer
FCN-8s = Fully convolutional version of VGG16 with 2 skip layer

Results

PASCAL VOC 11/12

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [5]	47.9	-	-
SDS [14]	52.6	51.6	$\sim$ 50 s
FCN-8s	**67.5**	**67.2**	**$\sim$ 100 ms**

NYUDv2

	pixel acc.	mean acc.	mean IU	f.w. IU
Gupta *et al.* [15]	60.3	-	28.6	47.0
FCN-32s RGB	61.8	44.7	31.6	46.0
FCN-32s RGB-D	62.1	44.8	31.7	46.3
FCN-32s HHA	58.3	35.7	25.2	41.7
FCN-32s RGB-HHA	**65.3**	**44.0**	**33.3**	**48.6**

SIFT Flow

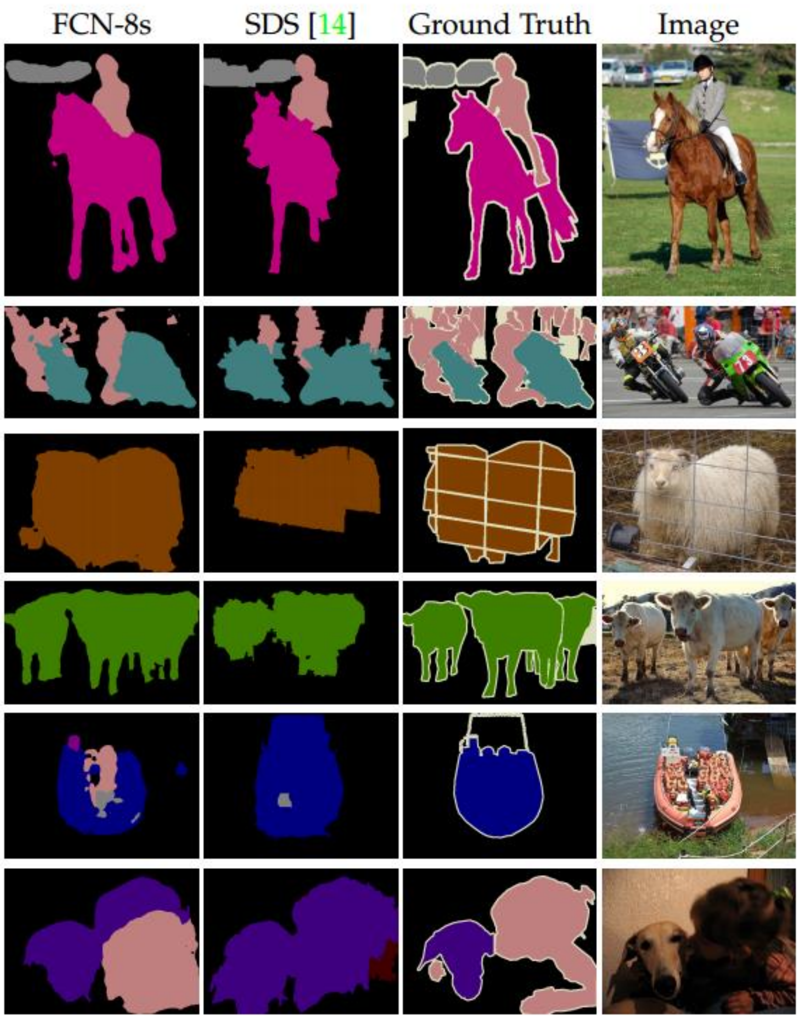
	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [57]	76.7	-	-	-	-
Tighe *et al.* [58] transfer	-	-	-	-	90.8
Tighe *et al.* [59] SVM	75.6	41.1	-	-	-
Tighe *et al.* [59] SVM+MRF	78.6	39.2	-	-	-
Farabet *et al.* [12] natural	72.3	50.8	-	-	-
Farabet *et al.* [12] balanced	78.5	29.6	-	-	-
Pinheiro *et al.* [13]	77.7	29.8	-	-	-
FCN-8s	**85.9**	**53.9**	41.2	77.2	**94.6**

PASCAL context

59 class	pixel acc.	mean acc.	mean IU	f.w. IU
O_2P	-	-	18.1	-
CFM	-	-	34.4	-
FCN-32s	65.5	49.1	36.7	50.9
FCN-16s	66.9	51.3	38.4	52.3
FCN-8s	**67.5**	**52.3**	**39.1**	**53.0**

FCN-32s = Fully convolutional version of VGG16
FCN-16s = Fully convolutional version of VGG16 with 1 skip layer
FCN-8s = Fully convolutional version of VGG16 with 2 skip layer

Results



FCN-32s = Fully convolutional version of VGG16
FCN-16s = Fully convolutional version of VGG16 with 1 skip layer
FCN-8s = Fully convolutional version of VGG16 with 2 skip layer

UNIVERSITY OF WATERLOO

Summary

- Contribution: Train FCNs end-to-end for pixelwise prediction
- Able to accept inputs of any size due to fully convolutional network
- Skip layers combines local and global features
- 30% improvement on PASCAL VOC2012, as well as improvement on other data sets

QUESTIONS