Fully Convolutional Network for Semantic Segmentation

Evan Shelhamer et al.

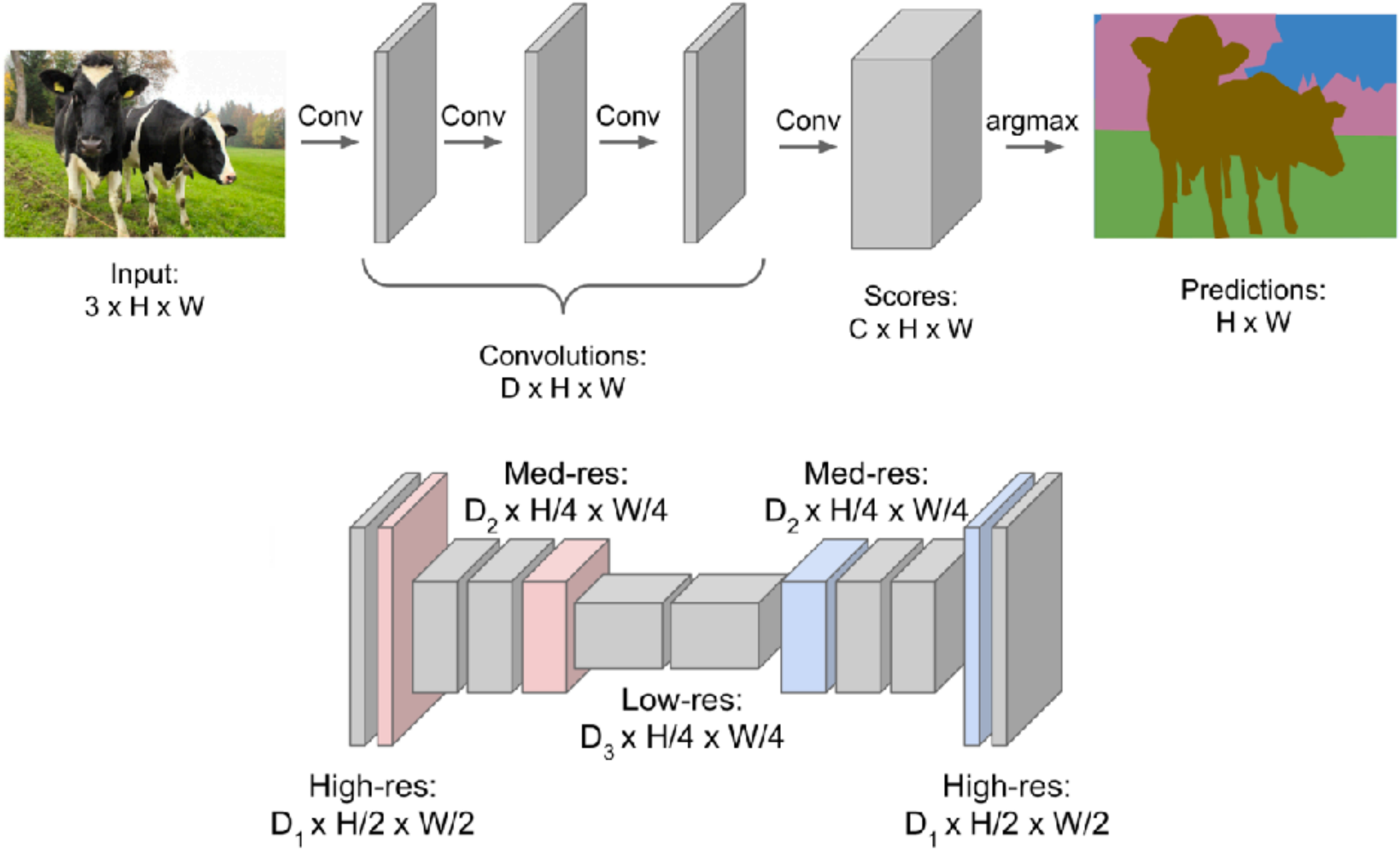
CVPR, 2015 &TPAMI, 2016

Kai Xie
kxie.cs@gmail.com

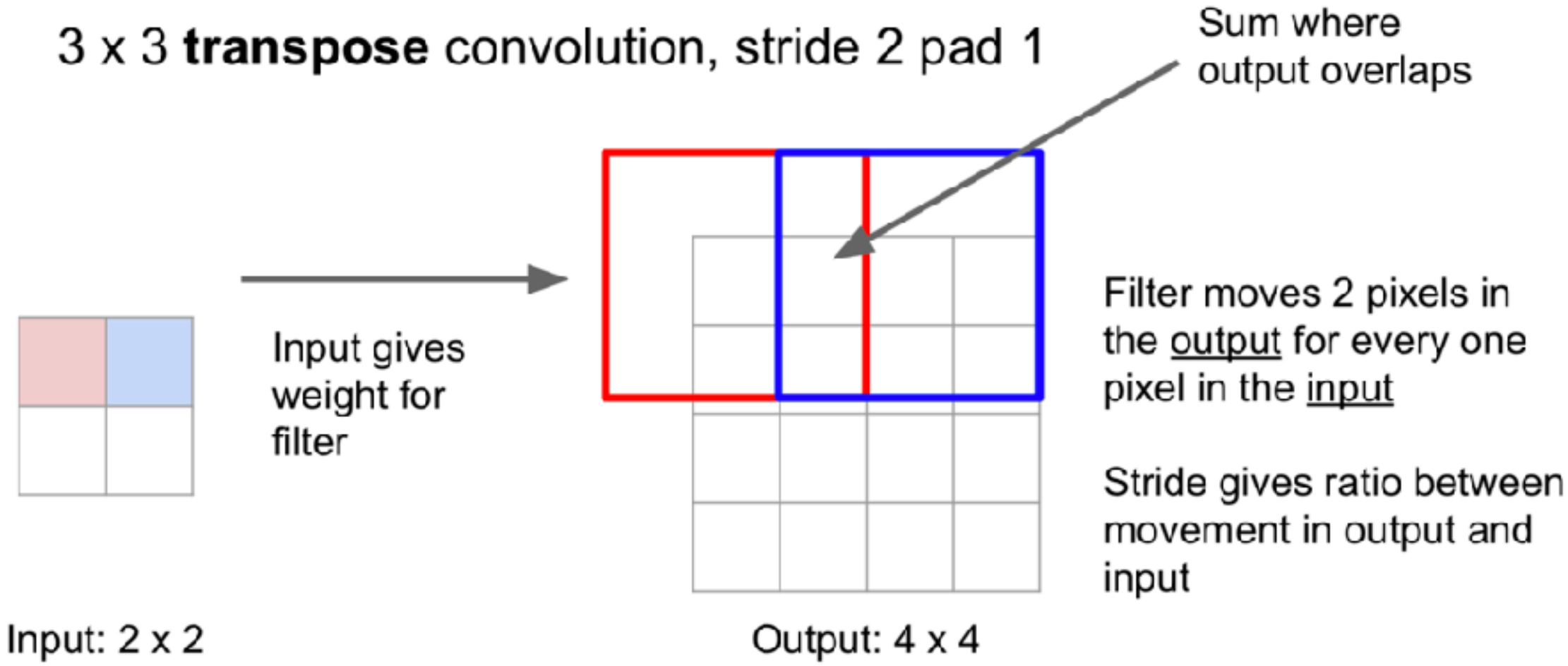
Outline

- Fully convolutional network
  - Main idea
  - Transpose convolution
  - Skip layers
- Results

Main Idea


Conv
Conv
Conv
Conv
argmax
Input:
3 x H x W
Convolutions:
D x H x W
Scores:
C x H x W
Predictions:
H x W
Med-res:
D2 x H/4 x W/4
Med-res:
D2 x H/4 x W/4
Low-res:
D3 x H/4 x W/4
High-res:
D1 x H/2 x W/2
High-res:
D1 x H/2 x W/2

Transpose Convolution



Skip layers

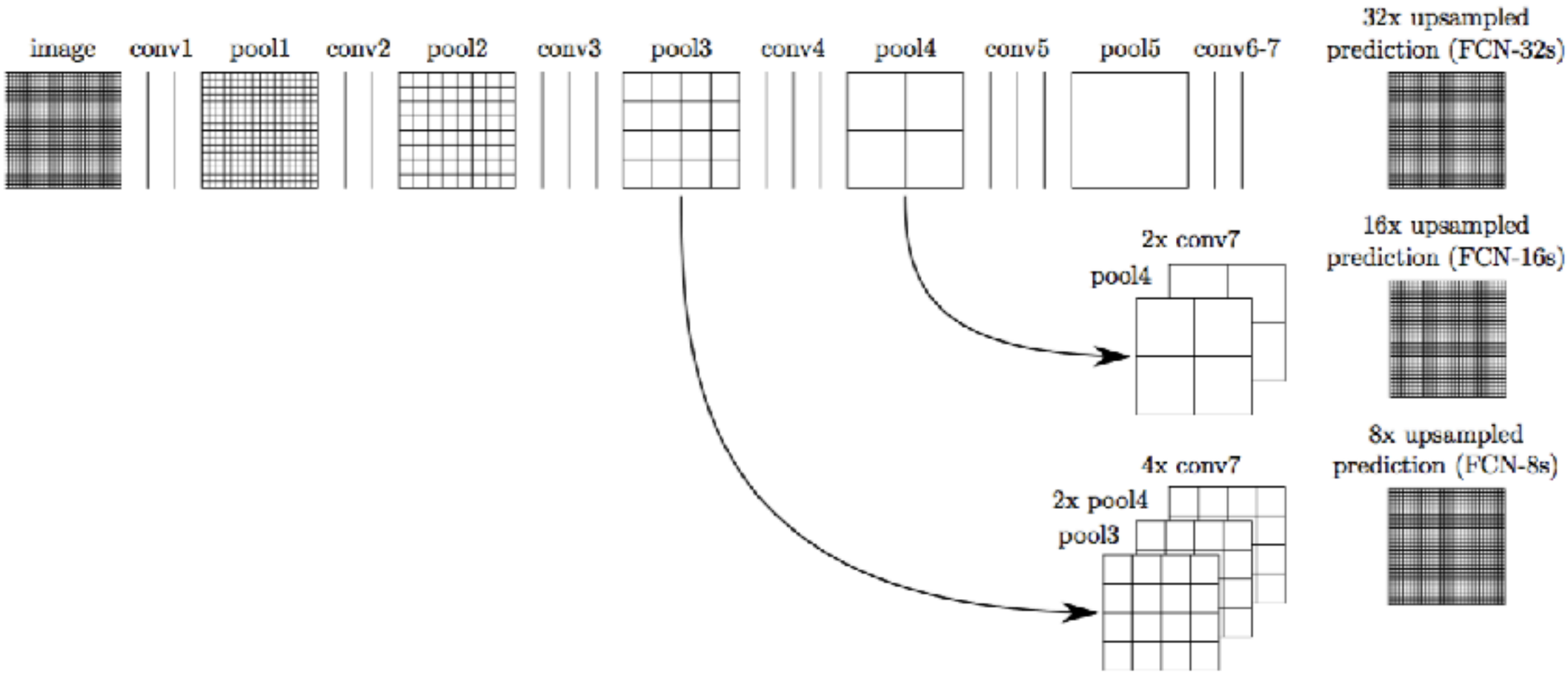


image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
32x upsampled
prediction (FCN-32s)
2x conv7
pool4
16x upsampled
prediction (FCN-16s)
4x conv7
2x pool4
pool3
8x upsampled
prediction (FCN-8s)

Results

Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{cl} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. We compute:

- pixel accuracy: $\sum_i n_{ii} / \sum_i t_i$
- mean accuraccy: $(1/n_{\text{cl}}) \sum_i n_{ii}/t_i$
- mean IU: $(1/n_{\text{cl}}) \sum_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$
- frequency weighted IU:
  $(\sum_k t_k)^{-1} \sum_i t_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$

Results

Table 2. Comparison of skip FCNs on a subset[7] of PASCAL VOC 2011 segval. Learning is end-to-end, except for FCN-32s-fixed, where only the last layer is fine-tuned. Note that FCN-32s is FCN-VGG16, renamed to highlight stride.

	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**

NYUDv2

Table 4. Results on NYUDv2. *RGBD* is early-fusion of the RGB and depth channels at the input. *HHA* is the depth embedding of [15] as horizontal disparity, height above ground, and the angle of the local surface normal with the inferred gravity direction. *RGB-HHA* is the jointly trained late fusion model that sums RGB and HHA predictions.

	pixel acc.	mean acc.	mean IU	f.w. IU
Gupta *et al.* [15]	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**

PASCAL VOC 2011 and 2012

Table 3. Our fully convolutional net gives a 20% relative improvement over the state-of-the-art on the PASCAL VOC 2011 and 2012 test sets and reduces inference time.

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [17]	52.6	51.6	~ 50 s
FCN-8s	**62.7**	**62.2**	~ **175 ms**

SIFT Flow

Table 5. Results on SIFT Flow[9] with class segmentation (center) and geometric segmentation (right). Tighe [36] is a non-parametric transfer method. Tighe 1 is an exemplar SVM while 2 is SVM + MRF. Farabet is a multi-scale convnet trained on class-balanced samples (1) or natural frequency samples (2). Pinheiro is a multi-scale, recurrent convnet, denoted RCNN$_3$ (o^3). The metric for geometry is pixel accuracy.

	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [25]	76.7	-	-	-	-
Tighe *et al.* [36]	-	-	-	-	90.8
Tighe *et al.* [37] 1	75.6	41.1	-	-	-
Tighe *et al.* [37] 2	78.6	39.2	-	-	-
Farabet *et al.* [9] 1	72.3	50.8	-	-	-
Farabet *et al.* [9] 2	78.5	29.6	-	-	-
Pinheiro *et al.* [31]	77.7	29.8	-	-	-
FCN-16s	**85.2**	**51.7**	39.5	76.1	**94.3**

Results

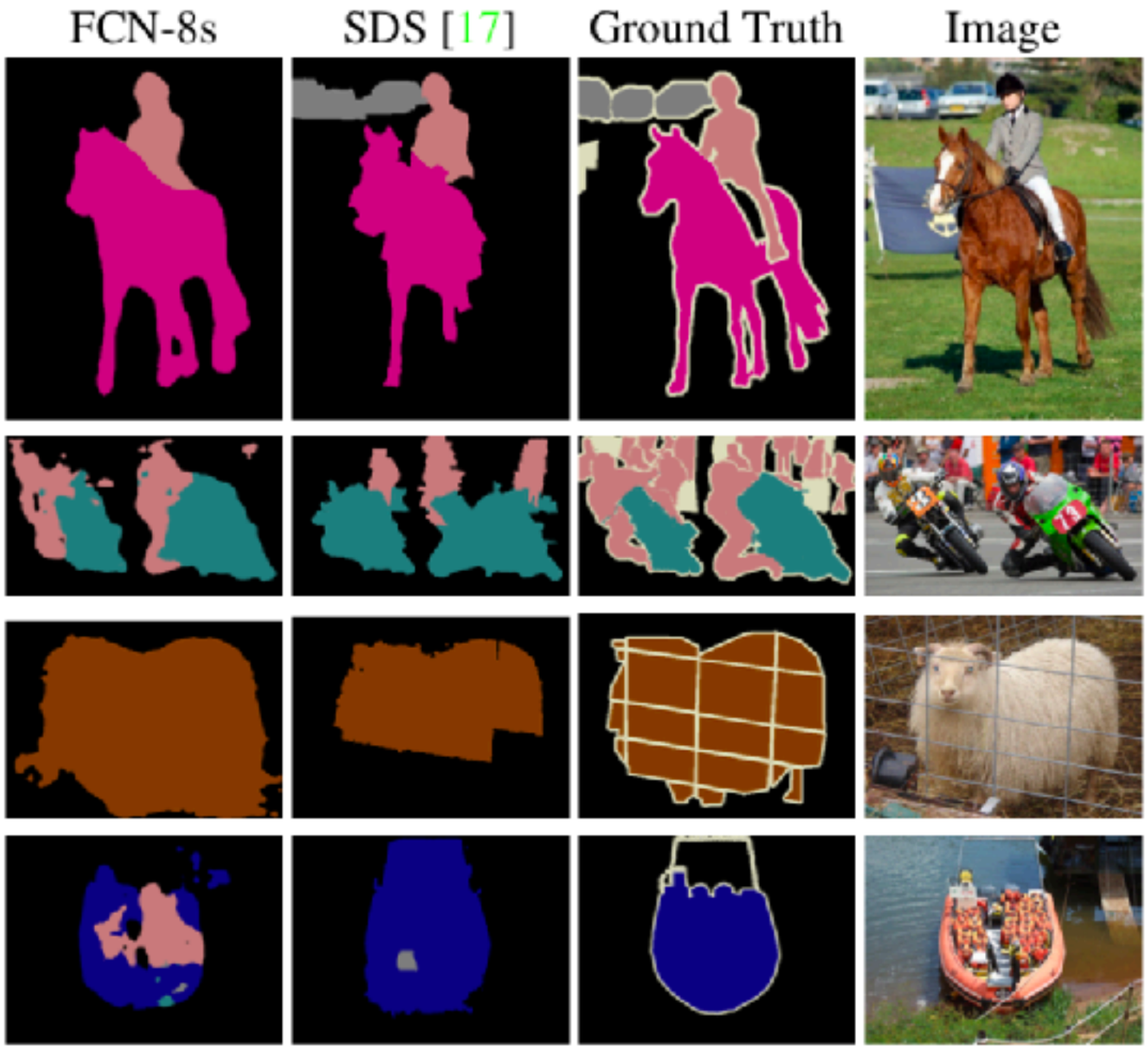


Figure 6. Fully convolutional segmentation nets produce state-of-the-art performance on PASCAL. The left column shows the output of our highest performing net, FCN-8s. The second shows the segmentations produced by the previous state-of-the-art system by Hariharan *et al*. [17]. Notice the fine structures recovered (first row), ability to separate closely interacting objects (second row), and robustness to occluders (third row). The fourth row shows a failure case: the net sees lifejackets in a boat as people.