Fully Convolutional Networks for Semantic Segmentation

Jonathan Long*, Evan Shelhamer*, Trevor Darrell
UC Berkeley

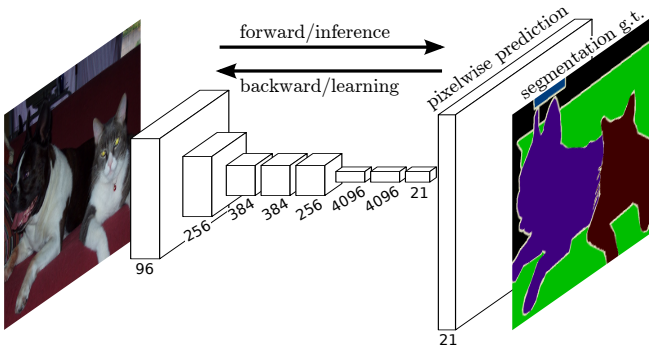

Figure 1: Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

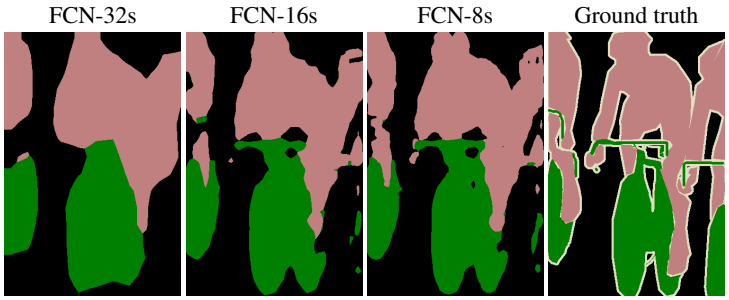

Figure 2: Refining fully convolutional nets by fusing information from layers with different strides improves segmentation detail. The first three images show the output from our 32, 16, and 8 pixel stride nets.

Fully convolutional networks (FCNs) trained end-to-end, pixels-to-pixels without further machinery exceed the state-of-the-art for semantic segmentation. Convolutional networks are powerful visual models that yield hierarchies of features. Our key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. We define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. We adapt contemporary classification networks (AlexNet [3], the VGG net [4], and GoogLeNet [5]) into fully convolutional networks and transfer their learned representations by fine-tuning [1] to the segmentation task. We then define a skip architecture that combines semantic information from deep, coarse layers with appearance information from shallow, fine layers to produce accurate and detailed segmentations. Our fully convolutional network achieves state-of-the-art segmentation of PASCAL VOC (20% relative improvement over SDS [2] to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes $\sim$ 175 ms for a typical image.

Semantic segmentation faces an inherent tension between semantics and location: global information resolves what while local information resolves where. Deep feature hierarchies encode location and semantics in a nonlinear local-to-global pyramid. To take advantage of this feature spectrum, we add skips between layers to combine multi-resolution information. We call this representation at a pixel the *deep jet*. This skip architecture is learned end-to-end to refine the semantics and spatial precision of the output.

FCN learning and inference are performed on a whole image at a time by dense feedforward computation and backpropagation. Fully convolutional versions of existing networks predict dense outputs from arbitrary-sized inputs. This lets us transfer the representations learned by classification models to semantic segmentation. In-network upsampling layers enable pixelwise prediction and learning in nets with subsampled pooling. To our knowledge, this is the first work to train FCNs end-to-end (1) for pixelwise prediction and (2) from supervised pre-training.

This method is efficient, both asymptotically and absolutely, and precludes the need for the machinery in other works. We compare to the common approach of patchwise training and find that our accelerated fully convolutional training does no harm to model quality. Our approach does not make use of pre- and post-processing complications including superpixels, proposals, image pyramids, post-hoc refinement by random fields or local classifiers, or ensembles.

We test our FCN on semantic segmentation and scene parsing, exploring PASCAL VOC, NYUDv2, and SIFT Flow. Although these tasks have historically distinguished between objects and regions, we treat both uniformly as pixel prediction. We evaluate our FCN skip architecture on each of these datasets, and then extend it to multi-modal input for NYUDv2 and multi-task prediction for the semantic and geometric labels of SIFT Flow.

Our code and models are publicly available at
http://fcn.berkeleyvision.org.

[1] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[2] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL http://arxiv.org/abs/1409.4842.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage. *Authors contributed equally.