Fully Convolutional Networks for Semantic Segmentation

Evan Shelhamer*, Jonathan Long*, and Trevor Darrell, *Member, IEEE*

Abstract—Convolutional networks are powerful visual models that yield hierarchies of features. We show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, improve on the previous best result in semantic segmentation. 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, the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. We then define a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Our fully convolutional network achieves improved segmentation of PASCAL VOC (30% relative improvement to 67.2% mean IU on 2012), NYUDv2, SIFT Flow, and PASCAL-Context, while inference takes one tenth of a second for a typical image.

Index Terms—Semantic Segmentation, Convolutional Networks, Deep Learning, Transfer Learning

1 Introduction

Convolutional networks are driving advances in recognition. Convnets are not only improving for whole-image classification [1], [2], [3], but also making progress on local tasks with structured output. These include advances in bounding box object detection [4], [5], [6], part and keypoint prediction [7], [8], and local correspondence [8], [9].

The natural next step in the progression from coarse to fine inference is to make a prediction at every pixel. Prior approaches have used convnets for semantic segmentation [10], [11], [12], [13], [14], [15], [16], in which each pixel is labeled with the class of its enclosing object or region, but with shortcomings that this work addresses.

We show that fully convolutional networks (FCNs) trained end-to-end, pixels-to-pixels on semantic segmentation exceed the previous best results without further machinery. To our knowledge, this is the first work to train FCNs end-to-end (1) for pixelwise prediction and (2) from supervised pre-training. Fully convolutional versions of existing networks predict dense outputs from arbitrary-sized inputs. Both learning and inference are performed whole-image-at-a-time by dense feedforward computation and backpropagation, as shown in Figure 1. In-network upsampling layers enable pixelwise prediction and learning in nets with subsampling.

This method is efficient, both asymptotically and absolutely, and precludes the need for the complications in other works. Patchwise training is common [10], [11], [12], [13], [16], but lacks the efficiency of fully convolutional training. Our approach does not make use of pre- and post-processing complications, including superpixels [12], [14], proposals [14], [15], or post-hoc refinement by random fields or local classifiers [12], [14]. Our model transfers recent success in classification [1], [2], [3] to dense prediction by reinterpreting classification nets as fully convolutional and fine-tuning from their learned representations. In contrast, previous works have applied small convnets without supervised pre-training [10], [12], [13].

Semantic segmentation faces an inherent tension between semantics and location: global information resolves *what* while local information resolves *where*. What can be done to navigate this spectrum from location to semantics? How can local decisions respect global structure? It is not immediately clear that deep networks for image classification yield representations sufficient for accurate, pixelwise recognition.

In the conference version of this paper [17], we cast pre-trained networks into fully convolutional form, and augment them with a skip architecture that takes advantage of the full feature spectrum. The skip architecture fuses the feature hierarchy to combine deep, coarse, semantic information and shallow, fine, appearance information (see Section 4.3 and Figure 3). In this light, deep feature hierarchies encode location and semantics in a nonlinear local-to-global pyramid.

This journal paper extends our earlier work [17] through further tuning, analysis, and more results. Alternative choices, ablations, and implementation details better cover the space of FCNs. Tuning optimization leads to more accurate networks and a means to learn skip architectures all-at-once instead of in stages. Experiments that mask foreground and background investigate the role of context and shape. Results on the object and scene labeling of PASCAL-Context reinforce merging object segmentation and scene parsing as unified pixelwise prediction.

In the next section, we review related work on deep classification nets, FCNs, recent approaches to semantic seg-

*Authors contributed equally

• *E. Shelhamer, J. Long, and T. Darrell are with the Department of Electrical Engineering and Computer Science (CS Division), UC Berkeley. E-mail: {shelhamer, jonlong, trevor}@cs.berkeley.edu.*

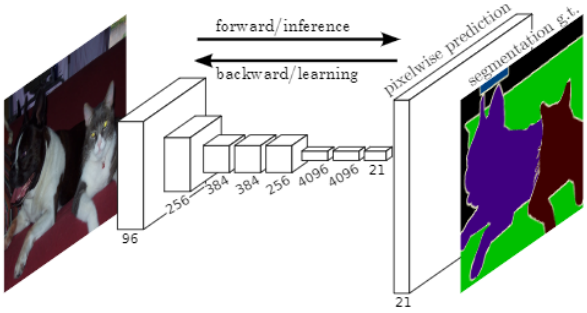

Fig. 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

mentation using convnets, and extensions to FCNs. The following sections explain FCN design, introduce our architecture with in-network upsampling and skip layers, and describe our experimental framework. Next, we demonstrate improved accuracy on PASCAL VOC 2011-2, NYUDv2, SIFT Flow, and PASCAL-Context. Finally, we analyze design choices, examine what cues can be learned by an FCN, and calculate recognition bounds for semantic segmentation.

2 Related Work

Our approach draws on recent successes of deep nets for image classification [1], [2], [3] and transfer learning [18], [19]. Transfer was first demonstrated on various visual recognition tasks [18], [19], then on detection, and on both instance and semantic segmentation in hybrid proposal-classifier models [5], [14], [15]. We now re-architect and fine-tune classification nets to direct, dense prediction of semantic segmentation. We chart the space of FCNs and relate prior models both historical and recent.

Fully convolutional networks To our knowledge, the idea of extending a convnet to arbitrary-sized inputs first appeared in Matan *et al*. [20], which extended the classic LeNet [21] to recognize strings of digits. Because their net was limited to one-dimensional input strings, Matan *et al*. used Viterbi decoding to obtain their outputs. Wolf and Platt [22] expand convnet outputs to 2-dimensional maps of detection scores for the four corners of postal address blocks. Both of these historical works do inference and learning fully convolutionally for detection. Ning *et al*. [10] define a convnet for coarse multiclass segmentation of *C. elegans* tissues with fully convolutional inference.

Fully convolutional computation has also been exploited in the present era of many-layered nets. Sliding window detection by Sermanet *et al*. [4], semantic segmentation by Pinheiro and Collobert [13], and image restoration by Eigen *et al*. [23] do fully convolutional inference. Fully convolutional training is rare, but used effectively by Tompson *et al*. [24] to learn an end-to-end part detector and spatial model for pose estimation, although they do not exposit on or analyze this method.

Dense prediction with convnets Several recent works have applied convnets to dense prediction problems, including semantic segmentation by Ning *et al*. [10], Farabet *et al*. [12], and Pinheiro and Collobert [13]; boundary prediction for electron microscopy by Ciresan *et al*. [11] and for natural images by a hybrid convnet/nearest neighbor model by Ganin and Lempitsky [16]; and image restoration and depth estimation by Eigen *et al*. [23], [25]. Common elements of these approaches include

- small models restricting capacity and receptive fields;
- patchwise training [10], [11], [12], [13], [16];
- refinement by superpixel projection, random field regularization, filtering, or local classification [11], [12], [16];
- "interlacing" to obtain dense output [4], [13], [16];
- multi-scale pyramid processing [12], [13], [16];
- saturating tanh nonlinearities [12], [13], [23]; and
- ensembles [11], [16],

whereas our method does without this machinery. However, we do study patchwise training (Section 3.4) and "shift-and-stitch" dense output (Section 3.2) from the perspective of FCNs. We also discuss in-network upsampling (Section 3.3), of which the fully connected prediction by Eigen *et al*. [25] is a special case.

Unlike these existing methods, we adapt and extend deep classification architectures, using image classification as supervised pre-training, and fine-tune fully convolutionally to learn simply and efficiently from whole image inputs and whole image ground thruths.

Hariharan *et al*. [14] and Gupta *et al*. [15] likewise adapt deep classification nets to semantic segmentation, but do so in hybrid proposal-classifier models. These approaches fine-tune an R-CNN system [5] by sampling bounding boxes and/or region proposals for detection, semantic segmentation, and instance segmentation. Neither method is learned end-to-end. They achieve the previous best segmentation results on PASCAL VOC and NYUDv2 respectively, so we directly compare our standalone, end-to-end FCN to their semantic segmentation results in Section 5.

Combining feature hierarchies We fuse features across layers to define a nonlinear local-to-global representation that we tune end-to-end. The Laplacian pyramid [26] is a classic multi-scale representation made of fixed smoothing and differencing. The jet of Koenderink and van Doorn [27] is a rich, local feature defined by compositions of partial derivatives. In the context of deep networks, Sermanet *et al*. [28] fuse intermediate layers but discard resolution in doing so. In contemporary work Hariharan *et al*. [29] and Mostajabi *et al*. [30] also fuse multiple layers but do not learn end-to-end and rely on fixed bottom-up grouping.

FCN extensions Following the conference version of this paper [17], FCNs have been extended to new tasks and data. Tasks include region proposals [31], contour detection [32], depth regression [33], optical flow [34], and weakly-supervised semantic segmentation [35], [36], [37], [38].

In addition, new works have improved the FCNs presented here to further advance the state-of-the-art in semantic segmentation. The DeepLab models [39] raise output resolution by dilated convolution and dense CRF inference. The joint CRFasRNN [40] model is an end-to-end integration of the CRF for further improvement. ParseNet [41] normalizes features for fusion and captures context with global pooling. The "deconvolutional network" approach of [42] restores resolution by proposals, stacks of learned deconvolution, and unpooling. U-Net [43] combines skip layers and learned deconvolution for pixel labeling of microscopy images. The dilation architecture of [44] makes