

Weakly- and Semi-Supervised Panoptic Segmentation

Qizhu Li , Anurag Arnab , and Philip H.S. Torr

University of Oxford

{liqizhu, aarnab, phst}@robots.ox.ac.uk

Abstract. We present a weakly supervised model that jointly performs both semantic- and instance-segmentation – a particularly relevant problem given the substantial cost of obtaining pixel-perfect annotation for these tasks. In contrast to many popular instance segmentation approaches based on object detectors, our method does not predict any overlapping instances. Moreover, we are able to segment both “thing” and “stuff” classes, and thus explain all the pixels in the image. “Thing” classes are weakly-supervised with bounding boxes, and “stuff” with image-level tags. We obtain state-of-the-art results on Pascal VOC, for both full and weak supervision (which achieves about 95% of fully-supervised performance). Furthermore, we present the first weakly-supervised results on Cityscapes for both semantic- and instance-segmentation. Finally, we use our weakly supervised framework to analyse the relationship between annotation quality and predictive performance, which is of interest to dataset creators.

Keywords: weak supervision, instance segmentation, semantic segmentation, scene understanding

1 Introduction

Convolutional Neural Networks (CNNs) excel at a wide array of image recognition tasks [1–3]. However, their ability to learn effective representations of images requires large amounts of labelled training data [4, 5]. Annotating training data is a particular bottleneck in the case of segmentation, where labelling each pixel in the image by hand is particularly time-consuming. This is illustrated by the Cityscapes dataset where finely annotating a single image took “more than 1.5h on average” [6]. In this paper, we address the problems of semantic- and instance-segmentation using only weak annotations in the form of bounding boxes and image-level tags. Bounding boxes take only 7 seconds to draw using the labelling method of [7], and image-level tags an average of 1 second per class [8]. Using only these weak annotations would correspond to a reduction factor of 30 in labelling a Cityscapes image which emphasises the importance of cost-effective, weak annotation strategies.

Our work differs from prior art on weakly-supervised segmentation [9–13] in two primary ways: Firstly, our model jointly produces semantic- and instance-segmentations of the image, whereas the aforementioned works only output instance-agnostic semantic segmentations. Secondly, we consider the segmentation of both “thing” and “stuff”

