Bottom-up Instance Segmentation using Deep Higher-Order CRFs

Anurag Arnab
anurag.arnab@eng.ox.ac.uk
Philip H.S. Torr
philip.torr@eng.ox.ac.uk

Department of Engineering Science
University of Oxford
United Kingdom

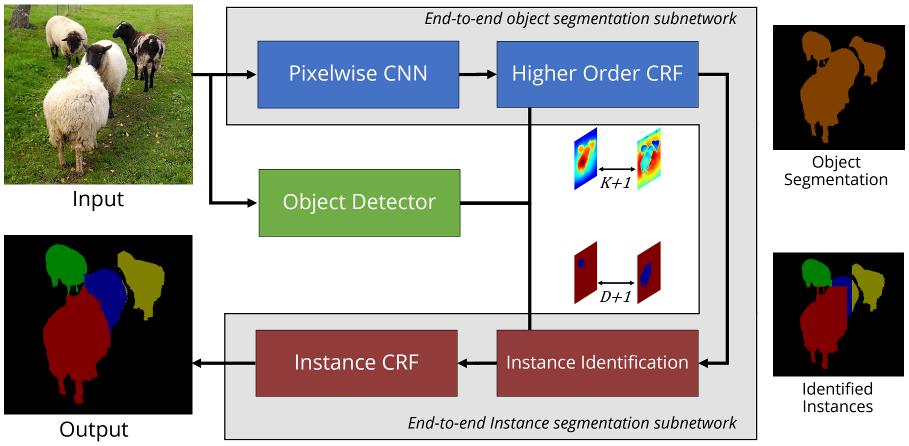

Figure 1. Overview of our end-to-end method. Our system first computes a category-level semantic segmentation of the image. A CRF with higher-order detection potentials is used to obtain this result in an end-to-end segmentation network. This results in a $W \times H \times (K+1)$ dimensional volume where K is the number of foreground classes. W and H are the image's width and height respectively. The original detections and the detector confidences recalibrated by the Higher Order CRF are used to identify instances in the image, producing a $W \times H \times (D+1)$ dimensional volume where D is the number of detections (variable per image). Thereafter, another Instance CRF is used to compute the final result.

Object detection and semantic segmentation have been two of the most popular Scene Understanding problems within the Computer Vision community. In this paper, we focus on the problem of *Instance Segmentation*. Instance Segmentation lies at the intersection of Object Detection – which localises different objects at a bounding box level, but does not segment them – and Semantic Segmentation – which determines the object-class label of each pixel in the image, but has no notion of different instances of the same class. As shown in Figure 1, the task of instance segmentation localises objects to a pixel level.

Many recent instance segmentation works have built on the "Simultaneous Detection and Segmentation" (SDS) approach of Hariharan *et al.* [2]. These methods all involve first detecting the various objects in an image before refining these detections into instance-level segmentations.

We present a different approach to instance segmentation, where we initially perform a category-level, semantic segmentation of the input image, classifying each pixel into one of K fixed categories. The resulting semantic segmentation is then refined into an instance-level segmentation, where the object class of each instance segment is obtained from the previous semantic segmentation. Both of these stages, while conceptually different, are fully differentiable and the entire system can be implemented as a neural network. We are able to reason about instances because our semantic segmentation network incorporates a differentiable Higher Order Conditional Random Field (CRF) which uses the cues from the output of an object detector. This CRF is inserted as another layer of a neural network [1, 4]. The object detection cues not only improve category-level segmentations, but the original detection scores are also calibrated during inference. This makes our system robust to false-positive detections, and helps us to reason about instances in the second-part of the network.

Assume an image $\mathbf{I}$ with N pixels, indexed $1, 2 \ldots N$, and define a set of random variables, $X_1, X_2, \ldots, X_N$, one for every pixel. We assign every pixel a label from a predefined set of object-category labels $\mathcal{L}$ such that each $X_i \in \mathcal{L}$. We also introduce latent binary random variables, $Y_1, Y_2 \ldots Y_D$ for every detection. If the d^{th} detection has been found to be valid after inference, Y_d will be set to 1, and 0 otherwise.

All latent Y_d variables are added to the CRF which previously only contained X_i variables. Let each $(\mathbf{X}_d, Y_d)$, where $\{\mathbf{X}_d\} = \{X_i \in \{\mathbf{X}\} | i \in F_d\}$, form a clique in the CRF. An assignment $(\mathbf{x}_d, y_d)$ to the clique $(\mathbf{X}_d, Y_d)$ has the energy:

$$\psi_d^{Det}(\mathbf{X}_d = \mathbf{x}_d, Y_d = y_d) = \begin{cases} w_l \frac{s_d}{|F_d|} \sum_{i=1}^{|F_d|} [x_d^{(i)} = l_d] & \text{if } y_d = 0, \\ w_l \frac{s_d}{|F_d|} \sum_{i=1}^{|F_d|} [x_d^{(i)} \neq l_d] & \text{if } y_d = 1, \end{cases} \tag{1}$$

where $x_d^{(i)}$ is the i^{th} element in vector $\mathbf{x}_d$, $[.]$ is the Iverson bracket and w_l is a class-specific, learnable weight parameter. This potential encourages consistency among $X_d^{(i)}$ variables and Y_d since it encourages $X_d^{(i)}$ variables to take the label l_d when Y_d is 1, and also encourages Y_d to be 0 when many $X_d^{(i)}$ variables do not take the label l_d. Note that this CRF contains usual unary and pairwise terms [3] as well.

Once we have a category-level segmentation of the image, each pixel still needs to be assigned to an object instance. We assume that each object detection represents a possible instance. Since there are D object detections (where D varies for every image), and some pixels do not belong to any object instance, but are part of the background, we have a labelling problem involving the labels, $\mathcal{D} = \{1, 2, \ldots D+1\}$

If a pixel falls within the bounding box B of a detection, we assign the pixel to that instance with a probability proportional to the rescored detection (obtained from the probability of the latent Y variable after inference) and the semantic segmentation confidence for that class.

$$\Pr(v_i = k) = \begin{cases} \frac{1}{Z(\mathbf{Y}, \mathbf{Q})} Q_i(l_k) \Pr(Y_k = 1) & \text{if } i \in B_k \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, v_i is a multinomial random variable indicating the "identified instance" at pixel i and takes on labels from $\mathcal{D}$, $Q_i(l)$ is the output of the initial category-level segmentation stage of our network and denotes the probability of pixel i taking the label $l \in \mathcal{L}$, and $Z(\mathbf{Y}, \mathbf{Q})$ is the normalisation factor.

This then acts as the unary potentials of another CRF with only unary and pairwise terms, which encourage appearance and spatial consistency [3]. These priors are valid in the case of instance segmentation as well. Although this final CRF deals with a different number of labels (instances) for every image, we are still able to formulate it in a differentiable way by using CRF parameters which are not class-specific as in [4].

Our simple, bottom-up method is able to effectively leverage the progress made by state-of-the-art semantic segmentation and object detection networks to perform the related task of instance segmentation. This is emphasised by our state-of-the-art performance on the VOC 2012 dataset.

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip H. S. Torr. Higher order potentials in end-to-end trainable conditional random fields. In *ECCV*, 2016.

[2] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. Springer, 2014.

[3] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.

[4] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.