Higher Order Conditional Random Fields in Deep Neural Networks

Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, Philip H.S. Torr

UNIVERSITY OF OXFORD

Aim

- End-to-end training of a Higher Order Conditional Random Field (CRF) for the problem of semantic segmentation.

Background

- Fully convolutional networks (FCNs) classify pixels independently of each other, and produce noisy predictions which do not respect image edges.
- To combat this, CRFs with pairwise terms [1], encouraging spatial and appearance consistency, are usually used as post-processing.

Our Approach with Higher Order Potentials

- We formulate a richer and more expressive CRF model which utilises two *Higher Order Potentials* (potentials defined over cliques of more than two variables).
- We use the *differentiable Mean Field inference* algorithm to obtain the most probable labelling, and incorporate it *as a layer of our neural network*.
- This allows *end-to-end training* of our Higher Order CRF with an FCN.
- *Detection potential* uses the complementary cues of an object detector to improve segmentations. It helps in cases where initial unaries are poor.
- *Superpixel potential* encourages consistency over larger regions, and removes spurious noise from the output.

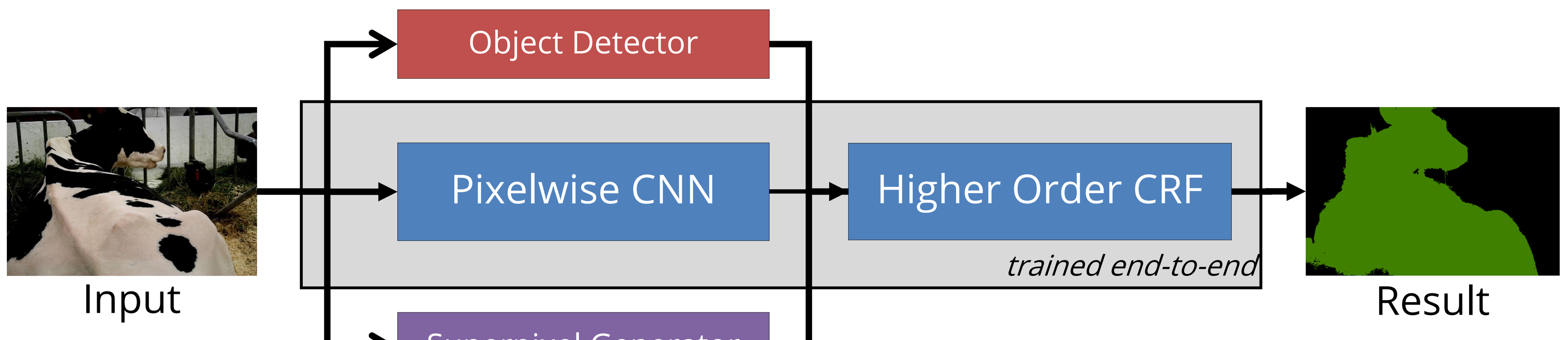

Figure 1: We train a Higher Order CRF end-to-end with a pixelwise CNN. Our higher orders improve significantly over only pairwise potentials [2].

Formulation

A Conditional Random Field is defined as

$$\Pr(\mathbf{X} = \mathbf{x}|\mathbf{I}) = (1/Z(\mathbf{I}))\exp(-E(\mathbf{x}|\mathbf{I}))$$

In our case, the energy (ignoring conditioning on Image $\mathbf{I}$) is:

$$E(\mathbf{x}) = \underbrace{\textstyle\sum_i \psi_i^U(x_i)}_{\text{Unaries from CNN}} + \underbrace{\textstyle\sum_{i<j} \psi_{ij}^P(x_i, x_j)}_{\text{Pairwise[1]}} + \underbrace{\textstyle\sum_d \psi_d^{\text{Det}}(\mathbf{x}_d)}_{\text{Detection potentials}} + \underbrace{\textstyle\sum_s \psi_s^{\text{SP}}(\mathbf{x}_s)}_{\text{Superpixel potentials}}$$

Detection Potential

Our detection uses the output of an object detector as additional cues for segmentation. Intuitively, object detectors can help when our pixelwise predictions are incorrect.

- Assume we have D object detections for a given image.
- The d^{th} detection is of the form (l_d, s_d, F_d)
  - $l_d \in \mathcal{L}$ is the class label of the d^{th} detection.
  - s_d is the detection score.
  - n_d is the number of foreground pixels in the d^{th} detection.
- Introduce binary latent variables, $Y_1, Y_2 \dots Y_D$ — one for each detection
  - Models whether detection is accepted or not.
  - $\Pr(Y_d = 1)$ initialised with s_d, the score of the object detector.
- $w_{\text{Det}}(l_d)$ is a learnable weight parameter that is a function of the class label.

This potential encourages consistency between detections, $\mathbf{Y}$, and labelled pixels, $\mathbf{X}$:

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

Superpixel Potential

Our learnable superpixel potential enforces consistency over regions obtained by superpixels. This is a soft constraint using a P^n-Potts type energy [4]. We use superpixels over multiple scales, which do not necessarily have to form a hierarchy.

$$\psi_s^{\text{SP}}(\mathbf{X}_s = \mathbf{x}_s) = \begin{cases} w_{\text{Low}}(l) & \text{if all } x_s^{(i)} = l, \\ w_{\text{High}} & \text{otherwise.} \end{cases}$$

Experimental Results on PASCAL VOC 2012

Table 1: Mean Intersection over Union (IoU) on the VOC Test Set compared to other works.

Method	Mean IoU [%]	Method	Mean IoU [%]
Ours			77.9
DPN [3]	77.5	Centrale [6]	75.7
Dilated [7]	75.3	BoxSup [8]	75.2
Attention [9]	75.1	CRF-RNN [2]	74.7

Table 2: Effect of each Higher Order potential on Reduced Validation Set.

Method	Mean IoU [%]
Baseline (Unary + Pairwise)	72.9
Superpixels Only	74.0
Detections Only	74.9
Superpixels and Detections	75.8

Figure 2: Output of system without superpixel potentials (left). Superpixels obtained from the method of [5]. Only one of the four "layers" is shown (middle). Note how the superpixel potentials get rid of spurious noise (right).

Base-line [2]

Ours

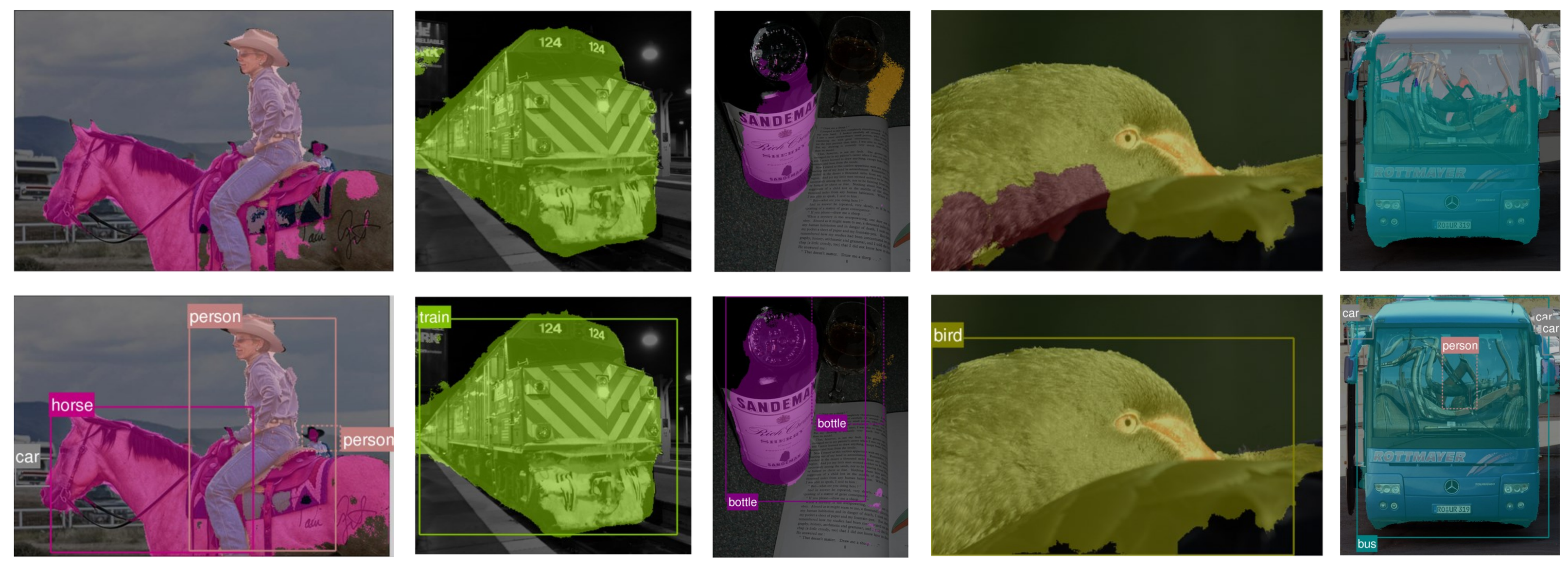

Figure 3: Qualitative comparison of our baseline with only pairwise potentials [2], and our method with higher orders. Our method uses object detection bounding boxes as an additional input, which are overlaid on the images.

Extension to Instance Segmentation

We have recently extended our detection potentials for the task of Instance Segmentation [10]. The detections inform us about possible object instances, and the problem is then to assign each pixel to an instance represented by a detection.

Figure 4: Instance Segmentation results using our Detection potential, as described in [10]. We produce both semantic segmentations (left) and instance segmentations

Conclusion

- Introduced two higher order potentials for a CRF which can be integrated into a deep neural network and trained end-to-end.
- Achieved the best performance on PASCAL VOC 2012 dataset at time of submission.
- In subsequent work [10], we have showed how our Detection potential can be used for the task of Instance Segmentation.

[1] P. Krahenbuhl *et al.* Efficient Inference in fully connected CRFs with Gaussian edge potentials. In *NIPS,* 2011.

[2] S. Zheng *et al.* Conditional Random Fields as Recurrent Neural Networks. In *ICCV* ,2015.

[3] Z. Liu *et al.* Semantic Segmentation via deep parsing network. In *ICCV,* 2015.

[4] P. Kohli *et al.* P3 & Beyond: Solving energies with higher order cliques. In *CVPR,* 2007.

[5] P. Felzenszwalb *et al.* Efficient graph-based image segmentation. *IJCV,* 2004.

[6] I. Kokkinos. Pushing the boundaries of boundary detection. In *ICLR,* 2016.

[7] F. Yu *et al.* Multi-scale context aggregation by dilated convolutions. In *ICLR,* 2016.

[8] J. Dai *et al.* Boxsup: Exploting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV,* 2015.

[9] L. Chen *et al.* Attention to scale: Scale-aware semantic image segmentation. In *CVPR,* 2016.

[10] A Arnab *et al.* Bottom-up Instance Segmentation with Higher Order CRFs. In *BMVC,* 2016

www.robots.ox.ac.uk/~tvg/projects/HoCrfCnn | aarnab@robots.ox.ac.uk