The Role of Context for Object Detection and Semantic Segmentation in the Wild

Roozbeh Mottaghi[1] Xianjie Chen[2] Xiaobai Liu[2] Nam-Gyu Cho[3] Seong-Whan Lee[3]
Sanja Fidler[4] Raquel Urtasun[4] Alan Yuille[2]
Stanford University[1] UCLA[2] Korea University[3] University of Toronto[4]
roozbeh@cs.stanford.edu, {cxj,lxb,yuille@stat}@ucla.edu,
{southq, swlee@image}@korea.ac.kr, {fidler,urtasun}@cs.toronto.edu

In this paper, we are interested in further analyzing the effect of context in detection and segmentation approaches. Towards this goal, we label every pixel of the training and validation sets of the PASCAL VOC 2010 `main` challenge with a semantic class (examples are shown in Figure 1). We selected PASCAL as our testbed as it has served as *the* benchmark for detection and segmentation in the community for years (over 600 citations and tens of teams competing in the challenges each year). Our analysis shows that our new dataset is much more challenging than existing ones (e.g., Barcelona [6], SUN [7], SIFT flow [5]), as it has higher class entropy, less pixels are labeled as "stuff" and instead belong to a wide variety of object categories beyond the 20 PASCAL object classes.

We analyze the ability of state-of-the-art methods [6, 1] to perform semantic segmentation of the most frequent classes, and show that approaches based on nearest neighbor retrieval are significantly outperformed by approaches based on bottom-up grouping, showing the variability of PASCAL images. We also study the performance of contextual models for object detection, and show that existing models have a hard time dealing with PASCAL imagery. In order to push forward the performance in this difficult scenario, we propose a novel deformable part-based model, which exploits both local context around each candidate detection as well as global context at the level of the scene. We show that the model significantly helps in detecting objects at all scales and is particularly effective at tiny objects as well as extra-large ones.

1. A Novel Contextual Dataset for PASCAL

We propose a dataset that contains pixel-wise labels for the 10,103 `trainval` images of the PASCAL VOC 2010 `main` challenge. There are 540 categories in the dataset, divided into three types: (i) objects, (ii) stuff and (iii) hybrids. *Objects* are classes that are defined by shape. This includes the original 20 PASCAL categories as well as classes such as fork, keyboard, and cup. *Stuff* denotes classes that do not have specific shape and appear as regions in images, e.g., sky, water. *Hybrid* classes are classes for which shape is so variable that it cannot be easily modeled, e.g., roads have clear boundaries (unlike sky), but their shape is more complex than the shape of a cup.

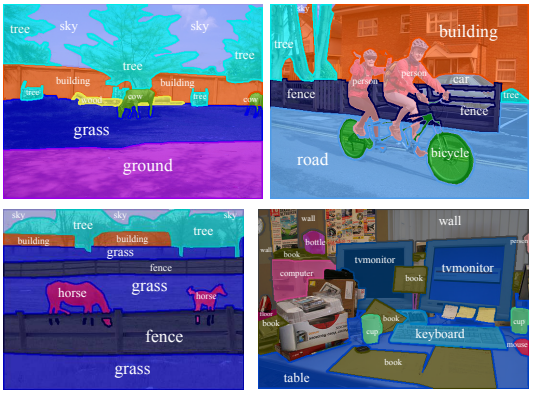

Figure 1. Examples of our annotations, which contain semantic segmentation of 540 categories in PASCAL VOC 2010.

2. A New Contextual Model

We designed a novel category level object detector, which exploits the global and local context around each candidate detection. By **global context** we mean the presence or absence of a class in the scene, while **local context** refers to the contextual classes that are present in the *vicinity* of the object. Following the success of [4], we exploit both appearance and semantic segmentation as potentials in our model. Our novel contextual model is a deformable part-based model with additional random variables denoting contextual parts, also deformable, which score the "contextual classes" around the object. Additionally, we incorporate global context by scoring context classes present in the full image. This allows us to bias which object detectors should be more likely to fire for a particular image (scene).

Unlike most existing approaches that *re-score* a set of

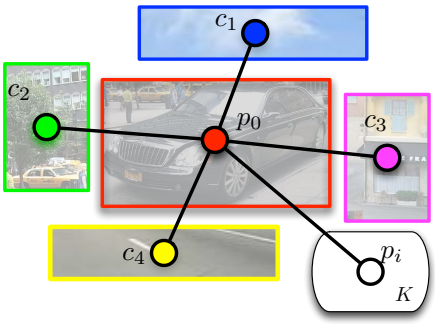

Figure 2. **Our model**: Context boxes are shown in color and correspond to top, bottom, left, and right boxes around the root filter.

boxes during post-processing, we perform contextual reasoning while considering exponentially many possible detections in each image. This is important as re-scoring-based approaches cannot recover from mistakes when the true object's bounding box does not appear among the set of detected boxes. An alternative is to reduce the detection threshold, but this will increase the number of false positives, lowering precision and increasing computation time.

The detection problem is framed as inference in a Markov Random Field (MRF), which scores each configuration of the root filter, as well as the two types of parts.

$$E(\mathbf{p},\mathbf{c}) = \underbrace{\sum_{i=0}^{K} \mathbf{w}_i^T \cdot \phi(x,p_i)}_{appearance} + \underbrace{\sum_{i=1}^{K} \mathbf{w}_{i,def}^T \cdot \phi(p_0,p_i)}_{part\ deformation} +$$

$$+ \underbrace{\sum_{j=1}^{C} \mathbf{w}_{j,lc}^T \phi(x,c_j)}_{local\ context} + \underbrace{\sum_{j=1}^{C} \mathbf{w}_{j,c.def}^T \phi(p_0,c_j)}_{context\ deformation} + \underbrace{\mathbf{w}_{gc}^T \phi_{gc}(x)}_{global\ context},$$

where x is the image, $\mathbf{c}$ is the set of contextual part placements and $\mathbf{p} = \{p_0, \cdots, p_K\}$, the root location, scale and component id, as well as the placements of the appearance parts. Fig. 2 illustrates the graphical model.

3. Contextual Segmentation Features

In order to decide on a particular segmentation algorithm to compute the features in our model we investigate two state-of-the-art algorithms: SuperParsing [6] and O2P [1] (applied to superpixels). We show the results of these methods on a few classes in Table 1. In general, this variation of O2P [1], which is based on bottom-up grouping outperforms SuperParsing [6], which is based on nearest-neighbor matching. So we choose O2P [1] to compute our contextual features.

	Recall		IOU	
	SuperParsing [6]	O_2P [1]	SuperParsing [6]	O_2P [1]
sky	88.8	95.1	83.0	87.1
water	44.4	74.6	42.4	67.9
grass	67.0	76.8	55.7	64.3
bus	23.0	71.7	23.8	58.1
tree	64.8	70.5	52.2	56.0
cat	37.1	70.2	32.7	53.5
aeroplane	29.6	67.2	30.6	52.6
motorbike	25.7	66.1	24.9	51.4
person	72.6	62.8	48.2	50.3
wall	65.8	73.1	46.1	48.9

Table 1. **Segmentation:** The results for 10 classes (out of 59 context classes) for which we obtain the highest accuracy.

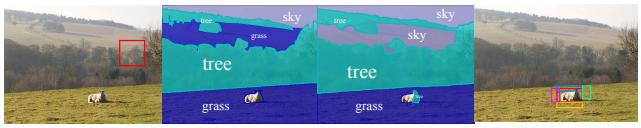

Figure 3. An example that is missed by DPM, but correctly localized when we incorporate context. We show the top detection of DPM, GT context labeling, context prediction by O_2P [1] and the result of our context model. Inferred context boxes are shown with different colors.

4. Object Detection and Segmentation in Context

We compare our method with [2]'s implementation of the Hierarchical Context model, and the context re-scoring method of [3], and show that our method better captures contextual information on PASCAL VOC 2010 `val` subset (30.8 mean AP vs. 26.7 and 27.8, respectively). An example detection is shown in Figure 3. We also show that a simple context feature can improve the performance of O2P [1], which has been the state-of-the-art on PASCAL segmentation in the past few years.

References

[1] J. Carreira, R. Caseiroa, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1, 2

[2] M. J. Choi, J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 2

[4] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for topdown detection. In *CVPR*, 2013. 1

[5] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. In *CVPR*, 2009. 1

[6] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 1, 2

[7] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1