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]

Abstract

In this paper we study the role of context in existing state-of-the-art detection and segmentation approaches. Towards this goal, we label every pixel of PASCAL VOC 2010 detection challenge with a semantic category. We believe this data will provide plenty of challenges to the community, as it contains 520 additional classes for semantic segmentation and object detection. Our analysis shows that nearest neighbor based approaches perform poorly on semantic segmentation of contextual classes, showing the variability of PASCAL imagery. Furthermore, improvements of existing contextual models for detection is rather modest. 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 this contextual reasoning significantly helps in detecting objects at all scales.

1. Introduction

Humans perceive the visual world effortlessly. We look at a complex and cluttered scene and know that the tiny object on the table is a fork and not the tail of an elephant. We know that the object hanging on the wall is more likely to be a picture or even a moose head than a car, and that a highly deformable entity stretching on the sofa is more likely to be a cat than a tiger. Context is a statistical property of the world we live in and provides critical information to help us solve perceptual inference tasks faster and more accurately.

Cognition-based studies have proved the effect of context in various perceptual tasks such as object detection, semantic segmentation and scene classification. The seminal work of Biederman et al. [3] and Hock et al. [17] showed that contextual information such as biases in object arrangements in particular scenes, relative physical size to other objects, and location are important cues for humans to detect objects. Furthermore, it is known that humans require a longer time to detect out of context objects. In a recent study, Parikh et al. [27] showed that context is an effective cue for humans to detect low-resolution (and typically small) objects in images. For object segmentation, Torralba [35] showed that at lower resolutions where only coarse scene information can be perceived, humans perform surprisingly well in delineating the most salient objects in the scene. In [26], the authors showed that humans are worse than machines at classifying small image patches but are far better when more contextual information is available.

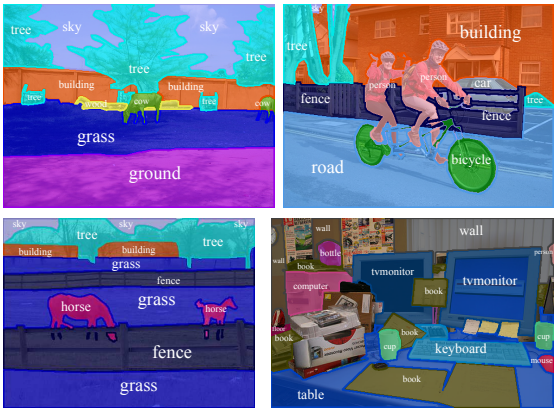


Figure 1. Examples of our annotations, which contain semantic segmentation of 540 categories in the PASCAL VOC 2010.

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 detection challenge with a semantic class. 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 [34], SUN [38], SIFT flow [25]), 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 [34, 7] to perform semantic segmentation of the most frequent classes, and show that approaches based on nearest neighbor (NN) retrieval are significantly outperformed by approaches based on bottom-up grouping, showing the vari-

ability 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.

2. Related Work

A number of approaches have employed contextual information in order to improve object detection [5, 28, 19, 36, 16, 9, 11]. This contextual information can be in the form of global scene context [36], ground plane estimation [28], geometric context in the form of 3D surface orientations [19], relative location [10], 3D layout [31, 14, 24], spatial support and geographic information [11]. In [16], contextual relationships between regions are found in an unsupervised manner and objects are detected using a discriminative approach. A context-driven search is proposed in [1] to focus on limited areas in the image to find the objects of interest. Torralba et al. [37] penalize the presence of objects in irrelevant scenes. In [9], spatial and co-occurrence priors are combined with local detector outputs and global image features to detect objects. The layout of familiar objects is used in [21] as context to find regions corresponding to unfamiliar objects. For both detection and segmentation, it has been shown that representing a region larger than an object itself leads to better performance [32, 6, 22, 8].

Holistic models that reason about the scene as a whole typically build strong contextual models to improve performance over tasks in isolation. [20, 40, 6] propose CRF models that reason jointly about object detection, image labeling and scene classification. In [29], the contextual consistency of inferred segment labels is imposed. In [23], improved performance is shown for scene classification when using a bank of object detectors instead of raw image features.

In recent years, a lot of effort has been invested in collecting densely labeled datasets. MSRC [32] was one of the first datasets with pixel-wise image labels, containing 592 images and 21 semantic classes. Camvid [4] contains 708 images of street scenes with 11 semantic classes. Recently, Liu et al. [25] released the SIFT flow dataset that contains 2688 images and 33 semantic labels, which are dominated by "stuff". SUN2012 [38], a subset of LabelMe, consists of 16873 images and 3819 object classes, most with only few training examples. Barcelona [34] is another subset of LabelMe, which includes 15150 images and 170 categories. Silberman et al. [33] released a dataset of indoor scenes containing 1449 RGB-D images and 894 object labels. The PASCAL VOC challenge has 11,530 training images containing 27,450 ROI annotated objects and 6,929 segmentations pertaining to 20 object classes. In this paper, we enrich these efforts, by labeling PASCAL VOC with pixel-wise accurate segmentation in terms of 520 additional classes.

3. A Novel Contextual Dataset for PASCAL

Our dataset contains pixel-wise labels for the 10,103 `trainval` images of the PASCAL VOC 2010 detection challenge (Fig. 1 shows example labels). 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.

Our annotation effort took three months of intense labeling performed by six in-house annotators. This resulted in much more accurate segmentations than when using online systems such as MTurk. While this increased the labeling cost significantly, we wanted to assure the highest possible accuracy and consistency of the annotations. The annotators were asked to draw a region and assign it a label using an interface similar to LabelMe [30]. There are about 12 regions in each image on average and the annotators spent about 3 to 5 minutes per image.

We provided the annotators with an initial set of 80 carefully chosen labels and asked them to include more classes if a region did not fit into any of these classes. Some cases were ambiguous to annotate; for example, the annotators were not sure how to label a tree visible through a window. We decided to go for a rich set of annotations, and thus allowed some pixels to have multiple labels (tree and window in this example). If the annotators were unable to recognize a region, they labeled it as unknown. We double checked each annotation and revised the ones that were not coherent in terms of category name or the region covering the object.

As expected the categories follow a power law distribution. For the analysis conducted in this paper, we select the 59 most frequent classes and assign to the rest the background label. As a consequence 87.2% of the pixels are labeled as foreground, and the rest as background. Note that the 20 object classes of PASCAL VOC cover only 29.3% of the pixels. Fig. 2 shows the distribution of pixels and images amongst these 59 most frequent categories.

Comparisons to existing contextual datasets: Several datasets exist that have been labeled with contextual classes. Notable examples are the Barcelona [34], SIFT flow [25] and SUN [38] datasets. We now show that our PASCAL-

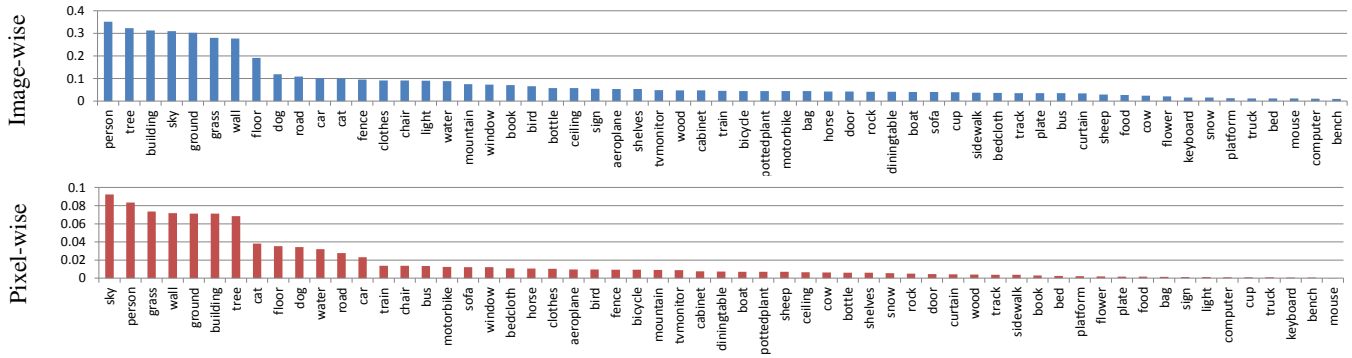


Figure 2. Distribution of pixels and images for the 59 most frequent categories. See text for the statistics.

context dataset has different statistics and is more challenging, and thus worth our efforts. Among the 35 most frequent categories of SUN [38], 87.2% of the pixels are "stuff", 94.5% for Barcelona [34], while 60.1% for PASCAL. Thus, the number of "things" and "stuff" pixels is more balanced in PASCAL. The entropy[1] for the most frequent 35 classes in Barcelona is 1.78, for SUN is 2.11 and for PASCAL is 3.11, which shows that more pixels are assigned to fewer classes in SUN and Barcelona. Thus, PASCAL images are more diverse than SUN's and Barcelona's.

Furthermore, PASCAL has served as *the* benchmark for detection and segmentation in the community for years. With the leaderboard made public just recently (`http://host.robots.ox.ac.uk:8080/leaderboard`), its popularity will even increase. Our annotations provide the community with dense labeling and 520 additional classes, thus giving plenty of information to exploit and new challenges to develop.

4. Object Detection In Context

In this section, we perform a detailed analysis of existing contextual models for detection in our PASCAL–CONTEXT dataset. Finally, we design a new contextual model, which is able to better exploit contextual information than existing approaches.

4.1. Existing contextual models

We explore two existing contextual models for object detection: DPM context re-scoring by Felzenswalb *et al*. [12] and Hierarchical Context by Choi *et al*. [9].

DPM Context re-scoring: The deformable parts-based model (DPM) [12] exploits contextual information in a simple way. Context re-scoring is used as a post processing step, which assigns a new score to each detected bounding box. The new score takes into account DPM scores of *all* other classes in an image, thus taking into account object co-occurrences in scenes, as well as the location and size of the box in order to exploit typical imaging biases. We make a slight modification by augmenting the original contextual features with the maximum confidence for 33 contextual categories, where the confidence of each such category is computed via a semantic segmentation method (details in Sec. 4.3).

Hierarchical Context: We investigated the contextual model of [9] which also re-scores DPM boxes. They assume a parent-child relationship between different objects and learn co-occurrence and spatial priors for objects related in the tree. Their goal is to infer which objects are present in the scene, the set of correct detections amongst candidates and their locations given global image features (i.e., GIST) and the output of DPM.

4.2. 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 [13], 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 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.

We follow the notation of [12], and define the root p_0 as a random variable encoding the location and scale of a

[1]For each category, we divide the number of pixels of that category to the total number of pixels in the dataset. This probability is used to compute entropy.

bounding box in an image pyramid as well as the mixture component id. This mixture is used to represent the appearance variability e.g., due to viewpoint. Let $\{p_i\}_{i=1,\cdots,K}$ be a set of *appearance parts* which encode part boxes at double the resolution of the root. Denote with $\{c_j\}_{j=1,\cdots,C}$ a set of variables describing the placement of our *contextual parts*. We model deformations between the root and both types of parts, penalizing the displacements with respect to an anchor position. These anchors represent the expected location of each part type with respect to the root. The appearance parts model discriminative/semantic parts of the object and are thus mostly inside the root's box, while the contextual parts model the typical surrounding of the object of a particular class, and are thus *outside* of the root's box. Following [12], we learn the anchors for the appearance parts from the training data. We use four contextual parts corresponding to the *top*, *bottom*, *left* and *right* side of the root filter. For the top/bottom parts, the height is set to $1/3$ of the height of the root filter and the width is the same as the root's. Fig. 3 illustrates the graphical model.

The detection problem is framed as inference in a Markov Random Field (MRF) [12], which scores each configuration of the root filter, as well as the two types of parts. We thus write the score of a configuration as the sum of four terms, appearance, context, and deformation:

$$E(\mathbf{p},\mathbf{c}) = E_{app}(x,\mathbf{p}) + E_{ctx}(x,\mathbf{c}) + E_{def}(\mathbf{p}) + E_{c.def}(\mathbf{c}),$$

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. Assuming a log-linear model, we define

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

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

As in [12], we use a HOG pyramid to compute $\phi(x,p_i)$. For both deformation costs we use the quadratic cost of [12].

We employ semantic segmentation to compute features for the contextual parts. In particular, we employ counts of pixels belonging to each contextual class inside each context part's box, normalized by the area of the part's box. We concatenate the normalized counts for all context classes to form our segmentation feature for each context part. As each feature is only summing within an area, it can be computed in constant time by employing a single integral image per context class. Note that our model is agnostic to the segmentation algorithm used. We explain our choice in 4.3.

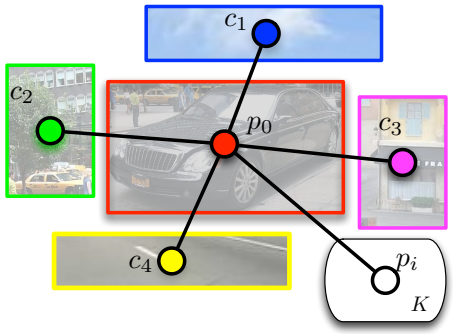


Figure 3. **Our model**: Context boxes are shown in color and correspond to top, bottom, left, and right boxes around the root filter.

For global context, we use a binary feature for each class, where 1 indicates that at least 1000 pixels were labeled with this class in the segmentation output. Note that global context does not depend on location, but on the class, and we learn a different weight for each class. We only use contextual classes and not also the object ones. This feature influences the detection score depending on global image information (i.e., type of scene): *e.g.*, if the image only contains pixels labeled as "sky" in the segmentation output, we are more likely to see a "plane" or "bird" and not a "tv".

Learning: We learn the model using latent structured SVMs. We utilize a 0-1 loss function based on IOU [15]. Stochastic gradient descent is used to optimize the non-convex objective. For initialization, we first train a mixture, root-only model as in [15], without context. We then add contextual parts, initializing the weights to 0, and perform several learning iterations of the model. We finally add the appearance parts, and train the full model using warm start.

Inference: Our model forms a tree, and thus exact inference is possible using dynamic programming. We start with the leaves, which require computing the score for each root filter, appearance and contextual parts. For the context parts, we first compute integral images for each contextual class, and then score a part in every location and scale of the pyramid. From here on, dynamic programming is agnostic about the type of part (appearance or context), thus computing the deformations and the final score as in [12].

4.3. 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 [34] and O2P [7].

SuperParsing [34] performs scene-level matching with the training set followed by superpixel matching. It then employs an MRF to incorporate neighboring contextual information. Performance is shown in Table 2. We employ IOU as well as recall as our metrics, where recall is just the percentage of correctly labeled pixels. We believe the main reason for the rather low performance is the high variability

	Recall	IOU		Recall	IOU
bag	1.4	1.3	**food**	16.3	14.9
bed	4.9	4.5	**mouse**	1.0	1.0
becloth	0.1	0.1	**plate**	11.5	9.6
bench	0.2	0.2	**platform**	10.5	10.2
book	11.7	9.6	**rock**	7.8	7.4
cabinet	6.9	6.3	**shelves**	14.9	10.2
clothes	4.3	3.9	**sidewalk**	0.5	0.5
computer	0.0	0.0	**sign**	10.7	9.9
cup	1.7	1.6	**snow**	18.6	17.0
curtain	21.9	19.1	**truck**	0.6	0.6
door	9.1	7.6	**window**	3.1	2.5
fence	12.0	10.0	**wood**	1.3	1.3
flower	13.2	12.5	**light**	14.6	12.4
			Avg.	7.6	6.7

Table 1. The subset of 59 most frequent classes that have low segmentation accuracy according to O_2P [7] results.

of PASCAL images, i.e., nearest-neighbor methods do not generalize well, requiring larger training sets.

O2P [7] uses shape-informed features to predict the amount of region's overlap with a GT segment for each class. The original method worked with bottom-up region proposals which were trained to detect object-like regions. Since our 59 classes of interest also include a large set of hybrid and stuff classes we decided for an alternative approach. In particular, we compute UCM superpixels [2], resulting in 67 superpixels per image on average. We then learn a classifier on the superpixels to predict their class using features based on SIFT, colorSIFT, and LBP as in [7].

As the output label of each superpixel we take the class with the highest confidence, and assign the background label if the scores of all foreground classes fall below a threshold, set to 0.35 empirically. As shown in Table 2, the performance of this approach is much higher than SuperParsing's. We thus choose to use it in our contextual detection model. In particular, accuracy for the "stuff" classes is very high (*e.g.*, sky has 87% IOU, water 68%, grass 65%). Some classes are, however, very difficult to segment (Table 1). Thus, to form our contextual features, we decided to not use classes in Tab. 1, and work only with 33 classes in Tab. 2.

4.4. Analysis of Contextual Detection Models

In this section, we analyze the results of the different contextual models. We used [9]'s implementation of the Hierarchical Context model, and the context re-scoring method of [12]. In both cases, we tuned their parameters to obtain the best possible performance. The training and evaluation have been performed on the `train` and `val` subsets of PASCAL VOC 2010 detection, respectively.

As shown in Table 3, our approach achieves the highest performance. For some classes our model is more effective, e.g., for bottle, sheep, or person the performance is 2-3% AP higher than the state-of-the-art method of [13], and provides 7-8% AP improvement for car and train. On the other hand, the performance for bicycle degrades, as the superpixels providing context cross the object boundary most of the time. Example detections are shown in Fig. 5.

	Recall		IOU	
	SuperParsing [34]	O_2P [7]	SuperParsing [34]	O_2P [7]
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
road	23.0	55.1	22.0	48.6
car	31.2	58.2	29.1	46.9
bicycle	16.5	55.6	16.2	44.6
keyboard	0.2	55.4	0.1	44.4
ground	48.9	51.9	38.7	41.7
floor	25.5	57.5	22.0	41.0
sheep	5.0	44.2	5.3	40.6
dog	18.8	46.9	17.5	39.9
bird	4.9	49.0	4.9	39.6
train	16.6	47.7	16.5	38.8
horse	2.2	44.9	2.1	38.8
tvmonitor	10.8	52.5	11.4	38.4
track	22.1	44.4	20.6	32.7
mountain	9.6	39.5	9.4	32.4
building	45.8	38.1	32.9	31.9
boat	0.9	37.5	1.1	31.7
pottedplant	1.1	35.9	1.2	29.4
sofa	4.4	33.6	5.5	28.2
table	9.4	33.7	7.0	27.7
bottle	1.3	35.4	1.4	27.7
ceiling	9.5	30.1	8.7	25.9
cow	0.1	25.0	0.1	24.0
chair	3.4	22.0	3.6	16.0
Avg.	25.1	52.2	21.7	42.4

Table 2. **Segmentation:** Nearest-neighbor methods such as [34] do not work well on PASCAL due to the high variability of images. In contrast the O_2P classifier [7] on superpixels performs well.

We tested [9] with our annotated contextual classes, and obtained marginal improvement (about 0.1 AP). To investigate the upper bound on the performance of this model, we used ground-truth (GT) information for context classes (excluding the 20 PASCAL classes) for both training and testing. Despite using GT, the improvement that [9] provides over DPM is quite marginal. We also evaluated DPM context rescoring [12] with our contextual classes. We performed the experiment in two settings with the original 20 object classes and also with 33 context classes. Adding more context classes does not help the method significantly.

We next analyze the effect of context when detecting different sizes of objects. We adopt the convention of [18] that cluster objects into XS (extra small), S (small), M (medium), L (large), and XL (extra large). As shown in Table 4, context improves detection for all sizes when using our method. The improvement is the largest for very small or very large objects. The appearance cues are weak for small objects, and they can be recognized mainly by their surrounding context (e.g., a boat that is far away from the camera has just a few pixels, but the water surrounding the boat provides a strong cue for detecting it). Very large objects are typically truncated making their detection hard for approaches that rely solely on appearance. This is not consistent with the claim of [11] that context is most useful only

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	mAP
DPM [12]	46.3	49.5	4.8	6.4	22.6	53.5	38.7	24.8	14.2	10.5	10.9	12.9	36.4	38.7	42.6	3.6	26.9	22.7	34.2	31.2	26.6
[9] + 33 context	46.2	40.5	10.5	12.4	16.5	53.2	36.2	28.2	17.2	8.3	14.5	18.8	27.9	35.4	41.9	9.3	28.7	19.9	38.6	28.6	26.7
[9] + GT 33 context	48.8	43.1	11.9	**14.3**	25.1	53.2	43.1	26.0	14.5	13.4	9.2	15.0	30.0	36.9	36.7	**10.9**	31.6	22.4	41.5	**35.5**	28.2
DPM rescoring [12] + 20 context	44.3	**51.3**	7.1	8.0	21.8	56.0	41.2	18.4	13.8	11.7	10.4	13.5	38.3	42.7	44.6	3.7	27.0	24.3	38.0	32.2	27.4
DPM rescoring [12] + 33 context	46.4	50.8	7.5	8.2	21.2	55.3	41.6	20.0	**14.7**	11.8	11.6	13.9	37.9	40.2	45.1	4.2	24.1	**27.6**	40.8	33.9	27.8
Ours + 20 context	46.9	50.1	9.2	9.5	**30.1**	**57.2**	44.1	30.7	12.7	**15.1**	12.9	14.2	35.6	**44.8**	44.0	4.9	30.6	20.1	42.2	34.8	29.5
Ours + 33 context	**49.8**	48.8	**12.0**	10.8	29.1	55.2	**45.6**	**32.0**	14.2	12.6	**13.7**	**16.6**	**39.8**	44.2	**45.1**	8.2	**35.3**	26.0	**42.3**	34.3	**30.8**

Table 3. Avg. Precision for detection of 20 PASCAL categories using 20 and 33 classes as context. The contextual information encoded by our contextual parts significantly outperforms the original DPM and other contextual models.

	XS	S	M	L	XL
DPM [12]	2.1	19.6	37.4	46.9	37.0
[9] + GT 33 context	4.6	22.7	36.1	44.2	34.5
DPM rescoring [12] + 33 context	2.5	20.0	38.5	48.0	41.0
Ours + 20 context	6.1	24.1	39.5	50.7	44.1
Ours + 33 context	7.6	25.8	41.1	53.1	46.8

Table 4. Effect of context as a function of size. Following [18], we have shown Normalized Average Precision across all categories to make the result of different size classes comparable.

for small objects in PASCAL. Another interesting observation is that although [9]+GT context performs better than DPM context rescoring, its performance for large objects is lower. Also, DPM rescoring has the most improvement for extra-large objects, but it is not very useful for small objects.

We also analyze how context affects false positives. For this purpose, we consider top 2000 false positive detections for each class and compute the confusion matrix for all approaches. We then subtracted the DPM confusion matrix from the confusion matrix of all contextual models. The results are shown in Fig. 4. There are some interesting trends. DPM context re-scoring is quite good at reducing the false positives from similar categories (e.g., cow-horse, bird-aeroplane, or car-bus), however, [9] and our method increase the confusion between objects that appear in similar context. Our method and [9] are mainly good at removing out-of-context objects, e.g., they reduce the confusion between cat and aeroplane, while DPM re-scoring increases it. Our method outperforms [9], for example, we reduce boat-aeroplane or train-bus confusion, but [9] increases it.

5. Object Segmentation In Context

We also evaluate the effect of context for *object* segmentation. Towards this goal, we augment the method of [7] with a simple feature which exploits the contextual classes in a region around the candidate bottom-up region. We choose [7] as it has been amongst the winners of PASCAL VOC segmentation challenge every year. It relies on a set of bottom-up object hypotheses which are computed by solving repetitively a figure/ground energy with different parameters and seeds via parametric min-cuts. This way of generating segments is called *CPMC*. These class independent hypotheses (segments) are then ranked based on a set of mid-level cues that capture general object characteristics. O_2P [7] performs segmentation by classifying and pasting high-ranked hypotheses into the image.

To integrate context with [7], for each CPMC segment, we consider a bounding box around the segment, which is $5/3$ times bigger than the tightest box around the segment along each dimension. For each bounding box, we remove the segment from it and extract the context feature on the rest of the pixels. Our contextual feature is a vector whose elements correspond to the maximum confidences from O_2P for the context classes depicted in Table 2. Our experiments show that adding a simple contextual feature produces nearly as much improvement as sophisticated methods that have been developed recently (e.g., [39]). The results are shown in Fig. 6 and Table 5. This is very encouraging, as we expect that encoding higher-level contextual information would provide even better performance.

6. Conclusions

In this paper, we studied the role of context in detection and segmentation approaches. Towards this goal, we labeled every pixel of the PASCAL VOC 2010 detection challenge. Our analysis showed that NN-type approaches perform very poorly in segmentation due to the variability of PASCAL imagery. Furthermore, improvements of existing contextual models for detection is rather modest. We have proposed simple ways to explore context in segmentation and detection and show their effectiveness. We expect our efforts to provide the community with plenty of new detection and segmentation challenges.

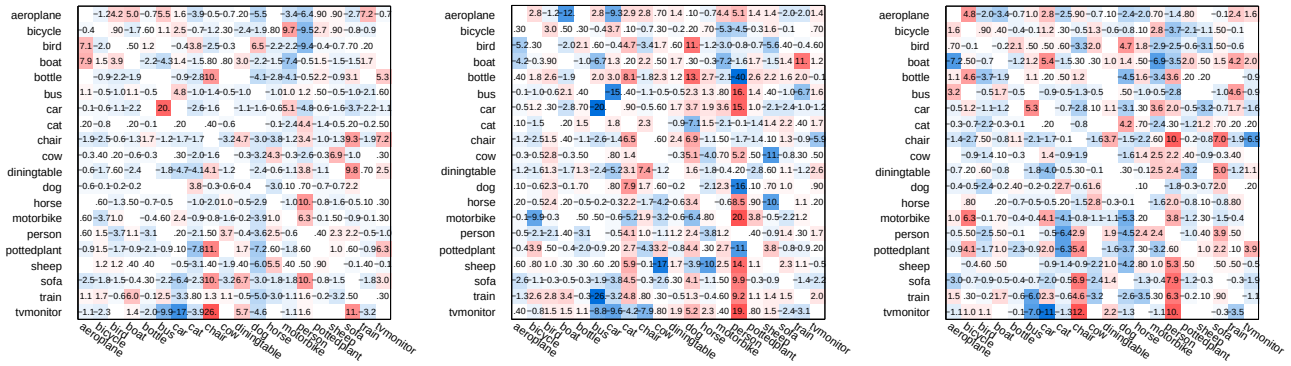
Acknowledgments: This work was supported by ARO 62250-CS, ONR N00012-1-0883 and ONR N00014-13-1-0721, and also by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Korean Ministry of Trade, Industry and Energy (Grant Number: 10041629).

References

[1] B. Alexe, N. Heess, Y. W. Teh, and V. Ferrari. Searching for objects driven by context. In *NIPS*, 2012. 2

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. In *PAMI*, 2011. 5

[3] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 1982. 1



(a) Hierarchical context [9] (b) DPM re-scoring [12] (c) Ours

Figure 4. Analysis of false positives in different contextual methods. The confusion matrices correspond to subtracting the DPM confusion matrix from the confusion matrix of each of the methods. To make a fair comparison, we consider top 2000 false positives for all methods.

Figure 5. Objects that are 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 and the result of our context model. Inferred context boxes are shown with different colors.

[4] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2008. 2

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004. 2

[6] G. Cardinal, X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010. 2

[7] J. Carreira, R. Caseiroa, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1, 4, 5, 6, 8

[8] J. Carreira, F. Li, and C. Sminchisescu. Object Recognition by Sequential Figure-Ground Ranking. *IJCV*, 2012. 2

[9] 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, 5, 6, 7

[10] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 2

[11] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 2, 5

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 3, 4, 5, 6, 7

[13] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for topdown detection. In *CVPR*, 2013. 3, 5

[14] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011. 2

[15] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *NIPS*, 2011. 4

[16] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 2

[17] H. Hock, G. P. Gordon, and R. Whitehurst. Contextual relations: The influence of familiarity, physical plausibility, and belongingness. In *Perception & Psychophysics*, 1974. 1

[18] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 5, 6

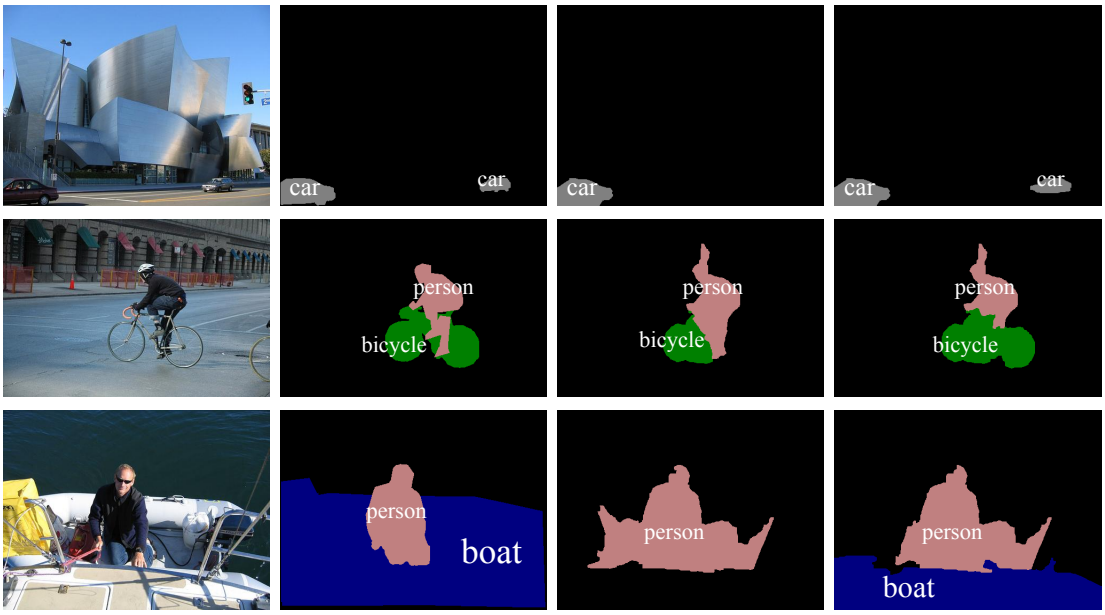


Figure 6. The result of augmenting O_2P [7] with a simple contextual feature.

	bg	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	Avg.
O_2P [7]	79.6	48.2	32.5	38.7	29.6	32.8	61.1	46.7	50.4	12.4	33.9	12.4	36.8	36.3	46.0	49.0	20.8	41.6	17.0	36.7	41.6	38.29
O_2P [7] + 13 context	79.4	52.4	32.8	40.1	33.1	34.4	60.5	47.8	50.2	12.8	32.8	13.0	36.3	36.9	44.5	48.6	20.1	41.8	16.7	40.1	40.7	38.83

Table 5. The result of object segmentation on PASCAL 2010 object detection subset. We augment [7] with a simple context feature. The improvement is in the same range as recent sophisticated methods such as [39].

[19] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005. 2

[20] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 2

[21] Y. J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *CVPR*, 2010. 2

[22] C. Li, D. Parikh, and T. Chen. Extracting adaptive contextual cues from unlabeled regions. In *ICCV*, 2011. 2

[23] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010. 2

[24] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013. 2

[25] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. In *CVPR*, 2009. 1, 2

[26] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh. Analyzing semantic segmentation using human-machine hybrid crfs. In *CVPR*, 2013. 1

[27] D. Parikh, C. L. Zitnick, and T. Chen. Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition. *PAMI*, 2011. 1

[28] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 2

[29] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 2

[30] B. Russell, A. Torralba, K. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 2008. 2

[31] A. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013. 2

[32] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, 2009. 2

[33] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[34] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 1, 2, 3, 4, 5

[35] A. Torralba. How many pixels make an image? *Visual Neuroscience*, 2009. 1

[36] A. Torralba, K. Murphy, and W. T. Freeman. Using the forest to see the trees: object recognition in context. *Comm. of the ACM*, 2010. 2

[37] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *CVPR*, 2003. 2

[38] 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, 2, 3

[39] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *CVPR*, 2013. 6, 8

[40] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2