Extracting Adaptive Contextual Cues from Unlabeled Regions

Congcong Li
Cornell University
cl758@cornell.edu

Devi Parikh
Toyota Technological Institute Chicago
dparikh@ttic.edu

Tsuhan Chen
Cornell University
tsuhan@ece.cornell.edu

Abstract

Existing approaches to contextual reasoning for enhanced object detection typically utilize other labeled categories in the images to provide contextual information. As a consequence, they inadvertently commit to the granularity of information implicit in the labels. Moreover, large portions of the images may not belong to any of the manually-chosen categories, and these unlabeled regions are typically neglected. In this paper, we overcome both these drawbacks and propose a contextual cue that exploits unlabeled regions in images. Our approach adaptively determines the granularity (scene, inter-object, intra-object, etc.) at which contextual information is captured. In order to extract the proposed contextual cue, we consider a scene to be a structured configuration of objects and regions; just as an object is a composition of parts. We thus learn our proposed "contextual meta-objects" using any off-the-shelf object detector, which makes our proposed cue widely accessible to the community. Our results show that incorporating our proposed cue provides a relative improvement of 12% over a state-of-the-art object detector on the challenging PASCAL dataset.

1. Introduction

Object recognition is one of the central problems in computer vision. Many recent works leverage contextual information surrounding the object-of-interest for enhanced recognition [5, 11, 12, 14, 19, 22]. These typically leverage labeled data from other object categories to learn contextual relationships. This leads to two undesirable consequences.

Unlabeled regions: First, regions in the images that are unlabeled are often neglected. In most scenarios such as the popular MSRC [1] or PASCAL [7] datasets, not all regions in the images are accounted for by the manually chosen categories that are labeled. For instance, 28.18% of the pixels in MSRC and 54.74% of pixels in the PASCAL 2007 dataset are not a part of the labeled categories. See Figure 1. Do these unlabeled regions contain useful information that is worth capturing? We conduct human studies on recognizing objects in low-resolution images[1] from the PASCAL dataset. We find that subjects perform better in scenarios where they see the entire image including the unlabeled regions (Figure 2). This indicates that the unlabeled regions, which are often discarded, indeed contain useful contextual information.

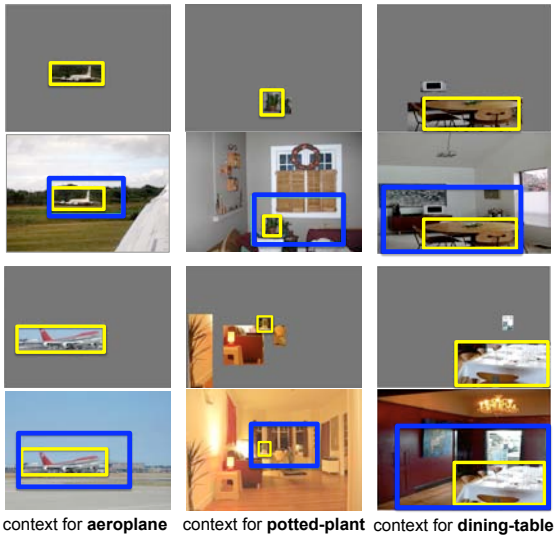


Figure 1. Many approaches to contextual reasoning for object detection model the relationship of the object-of-interest (yellow boxes) to other object categories labeled in the dataset. They do not leverage unlabeled regions in the images that do not belong to these manually chosen labeled categories, resulting in a highly myopic view of the scene (1^{st} and 3^{rd} row). Our approach (blue boxes) intelligently leverages information present in these unlabeled regions to better detect the object of interest. From left to right: unlabeled regions like sky and grass provide context for aeroplanes, and unlabeled objects like windows and paintings are contextually relevant for potted-plants and dining-tables respectively. The granularities of our blue boxes relative to yellow boxes are learned adaptively.

Adaptive granularity: Second, in focusing only on labeled regions in the image, models inadvertently limit the contextual information captured to the granularity implicit in the labels. For instance, most works exploring contextual models for the PASCAL dataset consider only object-level information, such as co-occurrence, relative location and relative size [5, 9]. We argue that the granularity at which contextual information is most helpful is category specific. While an "aeroplane" may only need to consider a

[1] Parikh *et al*. [19] showed that humans need contextual information for recognition only when the appearance information is impoverished, such as in low-resolution images.

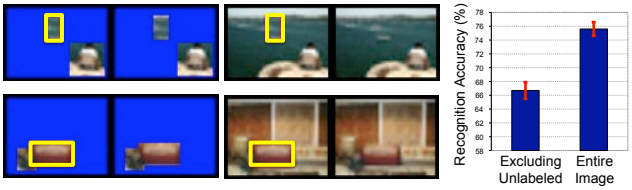


Figure 2. Human subjects were shown images excluding the unlabeled regions (left), as well as entire images (middle). The object to be recognized is shown with and without a yellow-outline (to avoid distraction). Our results (right) indicate that subjects can recognize objects significantly more reliably if information from the unlabeled regions is available. These experiments were conducted on 394 PASCAL 2007 images containing 897 objects from 20 categories.

neighboring sky region as context, "dining-table" can benefit from the entire scene layout. Moreover, while a "sheep" may be surrounded by relevant contextual information all around it, the most reliable contextual information for a "bicycle" is the person on top. Therefore, it is crucial that context is extracted from regions that adapt to different objects.

In this work, we overcome both these drawbacks and propose a contextual cue that exploits unlabeled regions in images and automatically adapts to each object category.

Formulation: What kind of information is contextually most useful? Information that is relevant *i.e*. has consistent spatial location with respect to the object-of-interest, and information that is reliable *i.e*. has consistent appearance across images for reliable detection. Interestingly, these are also aspects that make object models effective: capturing spatially coherent and visually consistent object-parts. Instantiating the popular hierarchical view of scenes [18, 20, 25] (parts are to objects as objects are to scenes), we can cast the problem of learning relevant contextual regions in a scene as that of learning detectors for "objects", which we call contextual meta-objects (CMOs). These CMOs form our proposed contextual cues. *Any* existing object detector can be used to learn our CMOs. In fact, the detector that one trains to detect objects-of-interest (OOIs) can be seamlessly (and conveniently so!) used to learn CMOs, essentially boosting the performance of the OOI detector without designing complex algorithms or cumbersome learning procedures.

Summary of approach: How can we extract meaningful contextual information from unlabeled regions? Our approach exploits the following key observation: object bounding boxes do not simply provide us with information about what the object looks like which can be used to train a detector for the object. The presence of an object at a certain location in a natural image also provides an anchor point that suggests a meaningful alignment among the scenes containing these objects. Given images labeled with bounding-boxes of the OOI, we align and cluster the images such that each cluster contains images with similar contextual information surrounding the OOI. We identify a CMO region around the OOI that best captures this contextual information. Note that the granularity at which the CMO region is defined is not fixed, and adapts to the content of the images. Any off-the-shelf object detector can then be trained to detect our CMOs. During testing, we apply the CMO detector on the test image, and the score of the detection captures our contextual cue. This cue can be combined with the OOI detection score, or other contextual cues for enhanced OOI detection.

Contributions: In this work, we effectively exploit the unlabeled regions in images that are often neglected by most existing works, to extract contextually relevant cues for enhanced object detection. Our contributions are threefold. First, we discover contextual regions that automatically adapt to the object category of interest in order to capture contextual interactions at varying granularities (the entire scene, inter-object, even intra-object) for different categories. Second, we cast the problem of extracting contextual regions in scenes into the problem of learning object models. This allows us to employ any off-the-shelf object detector to learn our contextual cue; a convenient choice being the object-of-interest detector whose performance we hope to enhance via contextual reasoning. This simplicity makes our proposed cue easily accessible to the community. Lastly, our approach achieves higher detection performance when using a *single* labeled category, than the state-of-the-art approach [9] that utilizes labels for *all* 20 object categories on the PASCAL 2007 dataset. This demonstrates our effective use of unlabeled regions. We use our approach to boost performance of several object detectors, and show that our contextual cue compares favorably to, and complements other sources of context. Our adaptive selection of the granularity of contextual information outperforms the fixed-granularity counterpart.

2. Related Work

Many recent works leverage contextual information for enhanced recognition and localization of objects in natural images. Various sources of context have been explored, ranging from the global scene layout, interactions between objects and regions, as well as local features. Divvala *et al*. [6] survey and study the effectiveness of different contextual cues and combine them to achieve superior performance. We view our novel cue that extracts useful contextual information from unlabeled regions as complementary to existing cues, and can be easily integrated in most contextual models.

Fixed-granularity models: Many existing works commit to a fixed granularity of contextual information. To incorporate scene-level information, Torralba et al. [26] use the statistics of low-level features across the entire scene to prime object detection. Hoiem et al. [13] use 3D scene information to provide priors on potential object locations. Park *et al*. [21] use the ground plane estimation as contextual cues for pedestrian detection. Probabilistic models have

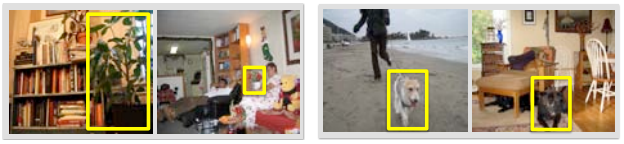
been proposed to capture the local interactions between neighboring regions [12, 14, 17], objects [5, 9, 19, 22, 30], or both [2, 10]. Sadeghi *et al*. [24] propose and detect visual phrases which correspond to chunks of meaning bigger than objects and smaller than scenes. While the visual phrases in [24] are manually labeled in the dataset, our work can be thought of as learning the visual phrases in an unsupervised way. Our learned composites, however, may not have a clearly defined semantic meaning. While most works focus on one level of interaction, Galleguillos *et al*. [10] explore contextual interactions at multiple levels. However, this multi-level aspect is gained by explicitly using different models for each interaction level. In contrast, our approach allows for adaptively picking different granularities of contextual information within the same framework.

Leveraging unlabeled information: A natural way to incorporate information from unlabeled regions in images is to build a global descriptor for the entire image to provide contextual cues. Though global image statistics show great potential in priming object detection [26], very modest improvement can be achieved when applied to datasets like PASCAL (also confirmed in our experiments), where images have poor alignments due to the high variance in object scales, poses, *etc*. [3]. Instead, local neighboring regions tend to be more effective. Wolf *et al*. [29] sample contextual information from pre-defined relative locations. Dalal *et al*. [4] show that simply increasing the size of the person bounding box by a small amount boosts the accuracy of pedestrian detection. This is equivalent to leveraging potentially unlabeled regions in close proximity of the object-of-interest. Our approach instead automatically and adaptively determines the extent of contextual information to be captured around different object categories. Felzenszwalb *et al*. [9] often detect parts that lie slightly outside the sliding window, and use these detections to refine their bounding box prediction. Lee *et al*. [15] utilize a few labeled categories to discover other object categories in the 'background' that have consistent appearance and are contextually coherent with the labeled categories. While similar in philosophy to our work, our solution is quite different, as is the problem setting of enhancing object detection.

3. Approach

In order to extract contextual cues that exploit unlabeled regions in images, we work in the most extreme setting where only one object category is labeled (with bounding-boxes) in the training dataset, leaving a large portion of the images unlabeled. We can seamlessly incorporate more labeled categories in our approach, as we describe later.

Our goal then is to extract useful contextual regions, given images labeled with ground-truth bounding-boxes for just one object category *i.e*. the object-of-interest (OOI). The ground-truth OOI may occupy different proportions of images across the dataset, as seen in Figure 3(a). Moreover, images may exhibit different contextual settings, as seen in Figure 3(b). To better model these, we first group the training images that exhibit similar contextual extent and content (Sections 3.1 and 3.2). As stated earlier, we cast our problem of extracting a contextual region in a scene as that of learning an object model, which we call contextual meta-object (CMO), for which any off-the-shelf detector can be used. Each cluster of images is used to train a different *component* of our CMO detector (Section 3.3). During testing, we run both the trained OOI and CMO detectors, and the score of the latter provides contextual information for the former. While any contextual reasoning model can be used to integrate the two, in our implementation we adopt the simple contextual re-scoring scheme from [9] (Section 3.4).



(a) (b)

Figure 3. Context around objects (potted-plant and dog shown in yellow boxes) can vary in extent (left) as well as content (right).

3.1. Contextual-extent-based clustering

We consider the contextual-extent of an image to be the portion of the image that lies outside of and surrounds the OOI. Intuitively, we wish to group all images that contain the OOI with similar poses, at similar scales, and in similar locations with respect to the rest of the scene.

We use all the training images, as well as their left-right flipped versions. To enhance consistency in the data, we first divide the images into 2 groups: one containing images with the OOI ground-truth bounding boxes on the right, and the other containing images with the OOI on the left. We then employ the following procedure for both groups. Consider an image I^i consisting of n_B ground-truth OOI bounding-boxes $\{B_j\}, j \in \{1, \ldots, n_B\}$. I^i is described by a set of n_B five-dimensional descriptors $F^i = \{f_1, \ldots, f_j, \ldots, f_{n_B}\}$

$$f_j = \left[\frac{|x^i_{tl} - \bar{x}^i_j|}{s^i_j} \; \frac{|y^i_{tl} - \bar{y}^i_j|}{s^i_j} \; \frac{|x^i_{br} - \bar{x}^i_j|}{s^i_j} \; \frac{|y^i_{br} - \bar{y}^i_j|}{s^i_j} \; \frac{h^i_j}{w^i_j} \right] \quad (1)$$

where $\left(x^i_{tl}, y^i_{tl}\right)$ and $\left(x^i_{br}, y^i_{br}\right)$ are the co-ordinates of the top-left and bottom-right corners of the image I^i respectively, $\left(\bar{x}^i_j, \bar{y}^i_j\right)$ are the co-ordinates of the center of the j^{th} ground-truth OOI bounding-box in image I^i, and s^i_j, h^i_j, and w^i_j are the scale, height and width of the j^{th} OOI bounding-box respectively. This descriptor captures the extent of the contextual information in the image (in terms of relative location and scale) with respect to the OOI present in the image, as well as the aspect ratio of the OOI. For each of the groups mentioned above, we cluster the descriptors $\cup_{i=1}^{N} F^i$ collected from all N training images containing the

OOI into K clusters using k-means clustering. This results in a total of $2K$ context-extent based clusters of images. In our implementation, we use $K = 3$.

For all images in the k^{th} cluster, $k \in \{1, \ldots, K\}$, we determine the extent of the CMO to be the largest bounding-box surrounding the OOI such that the CMO is entirely contained within at least 80% of the images in the cluster. This bounding-box indicates the presence and extent of the CMO we will learn via an off-the-shelf object detector. We note that since the extent of the CMO is not tied to the OOI bounding-box itself, and instead depends on the layout of the OOI in the scene, it can freely capture any relevant information in the image, including unlabeled regions, at any granularity. In fact, CMOs corresponding to the different clusters capture contextual information at different granularities for the same object category. We note that a training image I^i containing n_B ground-truth OOI instances will have n_B corresponding instances of the CMO for training.

3.2. Contextual-content-based clustering

While we ensure that the extent of context is similar among the images within the $2K$ clusters, the contextual setting or content of the images could be quite varied (Figure 3(b)). We further cluster each of the $2K$ groups based on their content. We extract gist [27] features within the CMO box, and perform k-means clustering on these features to divide each of the $2K$ groups into two clusters. In our implementation, when less than 40 images are assigned to a cluster, we drop the cluster and re-assign the corresponding images to the remaining clusters. Therefore, we have $M \leq 4K$ clusters, which varies with categories. For the PASCAL VOC 2007 dataset, we find a total of 198 clusters across the 20 categories. These clusters now have CMOs defined that are consistent in appearance as well as spatial relationships with respect to the OOI, and thus have the potential to provide useful contextual information to enhance the OOI detection.

3.3. Training CMO detectors

We now describe how we use the above clusters to learn our CMO detector. As stated earlier, we can use any off-the-shelf object detector, which we treat as a black-box parameterized by a model θ, learnt via a training procedure f_{train} that takes in training data of the form $(\mathcal{P}, \mathcal{N})$. $\mathcal{P} = \{(I_1^+, B_1), \ldots, (I_k^+, B_k)\}$ is a set of positive image and bounding-box pairs, and $\mathcal{N} = \{(I_1^-), \ldots, (I_l^-)\}$ is a set of negative images. The training procedure can be viewed as

$$\hat{\theta} = \text{optimize}\, f_{\text{train}}\left(\mathcal{P}, \mathcal{N}; \theta\right). \tag{2}$$

The detector can then be evaluated on a test image I, via the inference procedure f_{infer} to obtain a detection (B, s) including a bounding box B and a score s

$$(B, s) = f_{\text{infer}}\left(I; \hat{\theta}\right). \tag{3}$$

We see that the above formulation holds for a variety of detectors, be it sliding-window based [4, 9, 28] or hough-transform based [16]. So how do we use one of these black-box detectors to train our CMO detector using the M clusters formed in Section 3.2? Instead of committing to the hard-clustering which was blind to the choice of detector that would follow, we train an M-component detector using an EM style approach, similar to the strategy employed in [9]. We initialize the components using the above clustering, and train M detectors to obtain $\hat{\theta}^m$ using $f_{\text{train}}\left(\mathcal{P}^m, \mathcal{N}^m; \theta\right)$, where $\mathcal{P}^m$ is the set of images in the m^{th} cluster and the CMO windows contained (Section 3.1), and $\mathcal{N}^m$ are negative images. We then infer these M components on all positive training images (across all M clusters), and re-assign each "ground-truth" CMO box B (Section 3.1) in the training set to the component that best explains it:

$$\tilde{m} = \underset{m \in \{1, \ldots, M\}}{\text{argmax}} \left(s^m - t^m\right), \tag{4}$$

$$s.t. (B^m, s^m) = f_{\text{infer}}\left(I; \hat{\theta}^m\right), \text{overlap}(B^m, B) > 0.5 \tag{5}$$

where t^m is a bias term used to normalize the scores across the components, and is set to the 5^{th} percentile of the scores assigned by component m to all detections in the training images. Each component m is now re-trained, but with the new set of positive examples, to obtain an updated $\hat{\theta}^m$, and the iterations continue. We find that 3-5 iterations suffice. During testing, all M components are evaluated on the test image. Non-maximal suppression is used on the resultant detections to eliminate repeated detections.

3.4. Contextual re-scoring

We now describe how we use our CMO to provide context to the OOI detection. We evaluate our CMO as well as the OOI detectors on the test image. The presence as well as the location of the CMO can provide useful contextual information. However in this work, we only consider the presence *i.e.* the CMO detection score. Based on the CMO detections, we wish to re-score the detected OOI bounding boxes. While any contextual reasoning mechanism can be used to this end (such as [6, 9, 11, 23]), we train a classifier, similar to Felzenszwalb *et al.* [9]. Let D be the set of detections obtained for the OOI detector. Each detection (B, s), $(B, s) \in D$ is formed of a bounding-box with co-ordinates $B = (x_{tl}, y_{tl}, x_{br}, y_{br})$ (overloaded from previous sections) and a score s. Let s_C be the score of the highest scoring CMO detection (across all M components) found in the image. To re-score an OOI detection (B, s), we build a 6-dimensional descriptor consisting of the original score of the OOI detection, the top-left and bottom-right bounding box coordinates normalized by image size, and the contextual information provided by the CMO detection, captured via s_C:

$$\psi_{1c} = [\sigma(s)\ x_{tl}\ y_{tl}\ x_{br}\ y_{br}\ \sigma(s_C)] \tag{6}$$

where $\sigma(x) = 1/\left(1 + exp(-\alpha x)\right)$, $\alpha = 1.5$, as in [8].

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbik	pers	plant	sheep	sofa	train	tv	MEAN	GAIN
Base w/o context [9]	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3	–
Scene (∼ [26])	**30.9**	56.6	**11.5**	**18.5**	23.1	49.1	58.1	21.0	23.1	23.9	25.1	12.3	**59.9**	47.7	42.1	12.3	19.1	33.5	45.4	40.7	32.7	3.21%
EXO (∼ [4])	30.2	59.6	11.0	16.5	25.1	49.6	58.7	21.2	23.2	26.1	25.3	12.0	59.7	49.0	42.7	12.4	19.8	36.9	46.0	42.7	33.4	4.58%
CMO	30.5	**60.1**	11.2	17.0	**26.7**	**49.7**	**59.1**	**23.3**	**23.4**	**26.9**	**29.3**	**13.2**	59.7	**49.3**	**43.0**	**13.4**	**20.4**	**37.8**	**46.8**	**43.3**	**34.2**	**8.39%**

Table 1. Average precision (AP) for all 20 categories in PASCAL VOC 2007, mean AP across 20 categories, and the average relative improvement on 20 categories compared to the Base method. All methods listed use labels from only one object category. (∼[·]) means the method is similar in spirit to the reference work.

This ψ_{1c} descriptor is fed to a classifier h trained to separate correct OOI detections from false positives. The OOI bounding-box B is assigned a new score $\tilde{s} = h(\psi_{1c})$. In our implementation, the classifier h is an SVM with a polynomial kernel (parameters set via cross-validation), similar to [8, 9]. The training data are obtained by running the trained OOI detector on labeled training data, and collecting correct detections and false positives. We note that ψ_{1c} uses labeled data from only one object category, and still captures contextual information.

Our approach is not restricted to using labels from only one object category. If more object categories are available, they can be seamlessly incorporated. A CMO detector would be trained for each labeled object category. During test time, let $(D_1, \ldots, D_n)$ be the set of OOI detections obtained for n different object categories (for PASCAL $n = 20$ if all categories are considered). Let $(s_{C_1}, \ldots, s_{C_n})$ be the scores of the highest scoring CMO detections for each of the corresponding n CMO detectors. The contextual descriptor to re-score an OOI detection $(B, s) \in D_i$ is now $n + 5$ dimensional

$$\psi_{nc} = [\sigma(s)\ x_{tl}\ y_{tl}\ x_{br}\ y_{br}\ \sigma(s_{C_1}) \ldots \sigma(s_{C_n})]\,. \quad (7)$$

Similar to traditional approaches that exploit context, Felzenszwalb *et al*. [9] only use other OOI object categories to provide context, via a descriptor

$$\psi_{no} = [\sigma(s)\ x_{tl}\ y_{tl}\ x_{br}\ y_{br}\ \sigma(s_{D_1}) \ldots \sigma(s_{D_n})] \quad (8)$$

where s_{D_i} is the score of the highest-scoring OOI detection from the i^{th} OOI category. We note that this descriptor is identical to the one used in [9], which we compare to in our experiments.

Finally, a contextual descriptor capturing contextual information provided by all n CMO and OOI detectors is given as

$$\psi_{nco} = [\sigma(s)\ x_{tl}\ y_{tl}\ x_{br}\ y_{br}\ \gamma(C)\ \gamma(D)] \quad (9)$$
$$\gamma(C) = [\sigma(s_{C_1}) \ldots \sigma(s_{C_n})] \quad (10)$$
$$\gamma(D) = [\sigma(s_{D_1}) \ldots \sigma(s_{D_n})]. \quad (11)$$

A special case of Equation 9 for $n = 1$ differs from Equation 6 by one-dimension corresponding to $\sigma(s_D)$, the score corresponding to the highest-scoring OOI detection. Since the use of $\sigma(s_D)$ does not require additional training data, and is obtained by using labeled data from a single object-category, we replace Equation 6 with the following in our experiments.

$$\psi_{1co} = [\sigma(s)\ x_{tl}\ y_{tl}\ x_{br}\ y_{br}\ \sigma(s_C)\ \sigma(s_D)] \quad (12)$$

	CMO Ratio
OOI	37.2%
Other Categories	17.5%
Unlabeled Regions	45.3%

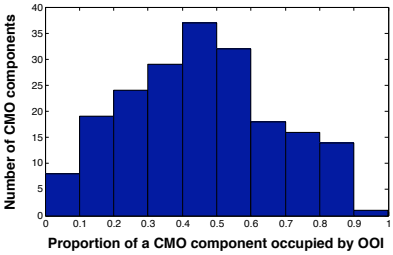


Figure 4. **Left**: The proportion of highest-scoring CMO detections on positive testing images occupied by different content. **Right**: The average proportion of a CMO component detection occupied by OOI. The truly multi-granular nature of our learnt contextual cues is evident.

4. Experiments and Results

We evaluate our approach using the PASCAL VOC 2007[2] challenge dataset and protocol [7], which contains 9963 images of realistic scenes, containing ground-truth bounding boxes for 20 object categories. While we provide results with other object detectors in subsequent experiments, we first perform several comparisons and analyses using the publicly available implementation [8] of the state-of-the-art deformable parts-based object detector [9].

4.1. Quantitative results

We first provide several quantitative evaluations, followed by some qualitative illustrations.

Useful contextual information from unlabeled regions: We first evaluate whether our proposed cue captures useful contextual information extracted from the unlabeled regions. We compare the performance of the baseline OOI detector to the contextually re-scored detector, using our proposed cue as the source of contextual information as described in Equation 12. The results can be seen in Table 1 (Base w/o context vs. CMO). We see that across all 20 categories, our method outperforms the state-of-the-art OOI detector, indicating that our contextual cue does in fact extract useful contextual information from unlabeled regions.

Further, in Figure 4 (left) we show the average proportion of the CMO detections occupied by different contents (OOI, other labeled categories, and unlabeled). We see that almost half of the CMO detections cover unlabeled regions. We also see that about 1/5 of the CMO detections capture

[2]We are not proposing a novel object detector. Instead, we will be performing a variety of comparisons to demonstrate the effectiveness of our proposed contextual cue. As recommended by the challenge organizers [7], we work with the 2007 test-set, which is the latest PASCAL test-set with publicly available annotations.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbik	pers	plant	sheep	sofa	train	tv	MEAN	GAIN
20OOI ([9])	31.2	61.5	11.9	17.4	27.0	49.1	59.6	23.1	23.0	26.3	24.9	12.9	60.1	51.0	43.2	13.4	18.8	36.2	**49.1**	43.0	34.1	7.57%
20CMO+20OOI	**31.5**	**61.8**	**12.4**	**18.1**	**27.7**	**51.5**	**59.8**	**24.8**	**23.7**	**27.2**	**30.7**	**13.7**	**60.5**	**51.1**	**43.6**	**14.2**	**19.6**	**38.5**	**49.1**	**44.3**	**35.2**	**12.05%**

Table 2. Average precision (AP) for all 20 categories in PASCAL VOC 2007, mean AP across 20 categories, and the average relative improvement on 20 categories compared to the Base method in Table 1). All methods listed use labels from all 20 categories.

# of labels for training	fusion methods	mean AP
1	Scene+EXO	33.6
	CMO	34.2
	Scene+EXO+CMO	34.4
20	20OOI	34.1
	20OOI+20CMO	35.2
	20OOI+Scene+EXO	34.6
	20OOI+20CMO+Scene+EXO	35.3

Table 3. The mean AP across the 20 categories in PASCAL VOC 2007 for fusing various sources of contextual information.

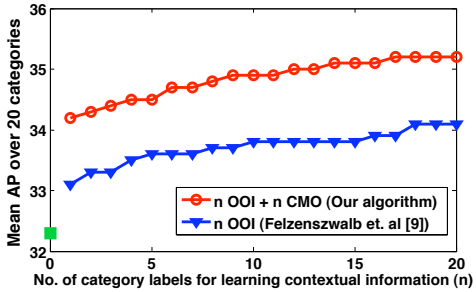


Figure 5. The effect of number of labeled categories on the average AP (across 20 categories).

other labeled categories in the images, even though our contextual cue does not use these labels to explicitly learn contextual relationships between the OOI and other categories.

Adaptive scaling helps: We now compare our contextual cue to two cues that also exploit the unlabeled regions but at a fixed granularity. The first cue ("Scene") captures the entire scene, and thus operates at the global granularity, similar to the work of Torralba et al. [26]. We train a binary RBF-kernel SVM classifier on the gist descriptors [27] extracted from the entire images to discriminate between images with and without the object-of-interest. We then use the same re-scoring scheme in Equation 12, by replacing the $\sigma(s_C)$ with $\sigma(s_g)$, where s_g is the score of the gist-based classifier for the test image. The second cue ("EXO") is local in nature. We expand the OOI bounding-box by a fixed amount (similar to [4]) of 20% in all four directions. Instead of training our CMO on the co-ordinates determined in Section 3.1, we use this expanded OOI bounding-box to learn a contextual "object" which we call EXO. We use the same re-scoring scheme in Equation 12, by replacing $\sigma(s_D)$ with $\sigma(s_E)$, where s_E is the highest score of the EXO detections. Note all three approaches only require labels from a single object category. Table 1 shows that our approach CMO outperforms both fixed-granularity methods. This demonstrates our ability to effectively determine the granularity of useful contextual interactions.

Further, Figure 4 (right) shows the distribution of the average proportion of CMO component detections occupied by OOI. High values (right of the histogram) correspond to contextual cues that capture local context around the OOI, while low values (left of the histogram) correspond to models that capture scene level context with respect to a relatively small OOI. The large variance in the distribution demonstrates the truly *multi-granular* nature of the contextual information learnt.

Complementary cue: We now test the ability of our proposed cue to provide complementary contextual information. We consider several different sources of context popularly explored in literature: the global scene-level context ("Scene") and local context ("EXO") described above, as well as object-level context provided by other labeled objects in the images, be it a subset of the categories (nOOI) or all 20 (20OOI). In Table 1 we saw that CMO performs better than the global and local context individually. Table 3 shows that our proposed cue learnt from a *single* labeled category is comparable to (slightly outperforms) the contextual information provided by *all* 20 labeled object categoires (20OOI). We note that 20OOI corresponds exactly to the contextual approach used by Felzenszwalb *et al.* [9]. Hence, we see that our contextual cue performs better than any of the individual sources of context, including ones that utilize significantly more amounts of labels.

Similar to [6], we analyze the performance of fusing various sources of contextual information. We use the same re-scoring method as described in Section 3.4 for combining the different cues, since we wish to evaluate the information captured by the cues, and not particular contextual reasoning techniques. Table 3 shows results of fusing different combinations of the above mentioned contextual cues. We append the highest score for each contextual cue being fused to re-score the OOI detection (similar to Equation 9). We see that our proposed cue individually performs better than the combination of *both* global and local context (Scene+EXO). Moreover, fusing our cue to these existing sources of context provides a further boost in performance, demonstrating the truly complementary nature of our cue that extracts contextual information from unlabeled regions at adaptive granularities.

We now compare our approach that leverages unlabeled regions to learn contextual cues (nOOI + nCMO, Equation 9), to the baseline approach of [9] (nOOI, Equation 8) that only utilizes other labeled categories, as the number of labeled categories is varied. For each object category (OOI), we pick n categories with highest mutual information (based on co-occurrence of categories with the OOI category across training data) to provide contextual information. Figure 5 shows the trends. The green point at $n = 0$

	OOI (mean AP)	OOI+CMO (mean AP)	Gain
ISM [16]*	12.6	15.6	23.8%
HOG-SVM [4]*	24.0	26.8	11.2%
Part-based Model [9]	32.3	34.2	8.4%

Table 4. Detection results with and without the proposed CMO as additional contextual information, by using different black-box detectors. $[\cdot]^*$ indicates it is our implementation of the reference work.

gives the mean AP of the OOI detector using no context. We can see that across the board, our approach leads to better AP than [9] while using the same amount of labeled data, demonstrating our effective use of unlabeled regions.

A break-down of the average AP across categories when using all 20 labeled categories for both methods can be seen in Table 2. We see that incorporating our contextual cue that leverages unlabeled information in addition to the 20 labeled object categories (20OOI+20CMO) provides improvements over 20OOI ([9]) in 19 out of 20 categories and matches the remaining category. We see that our contextual cue on average provides a relative improvement of 12.05% over the state-of-the-art detector. We observe significant relative improvements of more than 20% in some categories such as bird, cat, dining-table, and dog.

Other detectors: As mentioned in Section 3, our proposed contextual cue can be extracted via any object detector, and can inturn be used to enhance the performance of the detector for the OOI. We demonstrate that here. In addition to the sliding window part-based deformable model (Parts-based Model) by Felzenszwalb *et al*. [9] that we use in the above experiments, we consider two other popular detectors: Implicit Shape Model (ISM) [16] which is hough-transform based, and the HOG-SVM detector [4] (also sliding window). These are used as the black-box modules to train our CMO detectors, using the procedure described in Section 3.3. Table 4 shows the mean AP across all 20 categories in the PASCAL 2007 dataset, achieved by our implementation of these detectors, with and without the context provided by CMO. The results show that our proposed contextual cue CMO can be learnt via any detector, and consistently improves the object detection performance with a significant gain. Note that we use the same detectors for both OOI and CMO, indicating that no additional techniques are required for achieving these performance gains.

4.2. Qualitative results

Figure 6 shows some example CMO detections using the deformable parts-based model [9] as the detector. As quantitatively demonstrated earlier, we see that the CMO bounding-boxes contain a lot of unlabeled regions that are not labeled in the dataset, such as the sky region for aeroplanes, road for bicycle, coffee table and wall paintings for sofa, windows for potted-plants, *etc*. Although our approach uses labeled data only from one object category to learn the contextual cues, we learn meaningful relationships among objects as well as object parts. For example,

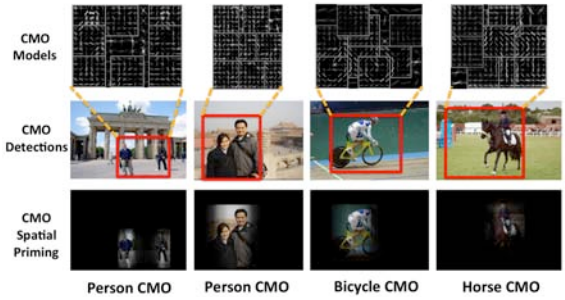


Figure 7. The spatial maps indicate the likelihood of a person being present given the CMO detections (red boxes). We see that the CMO provides strong priming for locations of OOIs.

the CMO detections for bicycle in the 2^{nd} column of Figure 6 consistently include a person's body as 'parts' of the CMO. In the 5^{th} and 6^{th} columns, we show detections of two CMO components corresponding to the person category. We see that both components seem to capture two people, but while the left-column models two people further apart from each other, the right-column detects two people close together. We also see intra-object CMO for cat in the last column. The false-positives shown in Figure 6 clearly demonstrate that our learnt CMO models fire at contextually relevant regions in images. Finally, as seen in Figure 4, these examples qualitatively demonstrate that our cues can adaptively learn contextual information at different granularities in the scenes.

Encouraged by the meaningful spatial interactions observed in Figure 6 (especially for the person category), we elaborate on that aspect a little further. As we demonstrated earlier, our learnt CMO models often include objects from other labeled categories (*e.g*. the bicycle CMO often includes a person, a person CMO often includes another person, *etc*.). By examining the CMO detections on training images, we can learn a distribution of the location of each labeled category in the dataset relative to our CMO models. These spatial distributions can now be used as a prior to better guide the detection of an OOI. We show a few examples of these spatial maps in Figure 7. We also display the HOG feature visualizations for the corresponding CMO. While in this work we only leverage CMOs to provide co-occurrence based context, exploring the potential of our models to capture scene configurations and provide explicit spatial priming for localizing objects is part of future work.

5. Conclusion

Regions that are not accounted for in the manually chosen list of categories labeled in a dataset are often neglected by existing works on contextual reasoning. In this work, we exploit these unlabeled regions to extract adaptive contextual cues for enhanced object detection. We utilize the labeled object bounding-box as an anchor to align scenes and learn spatially consistent and visually identifiable contextual regions. The granularity of these regions is adap-

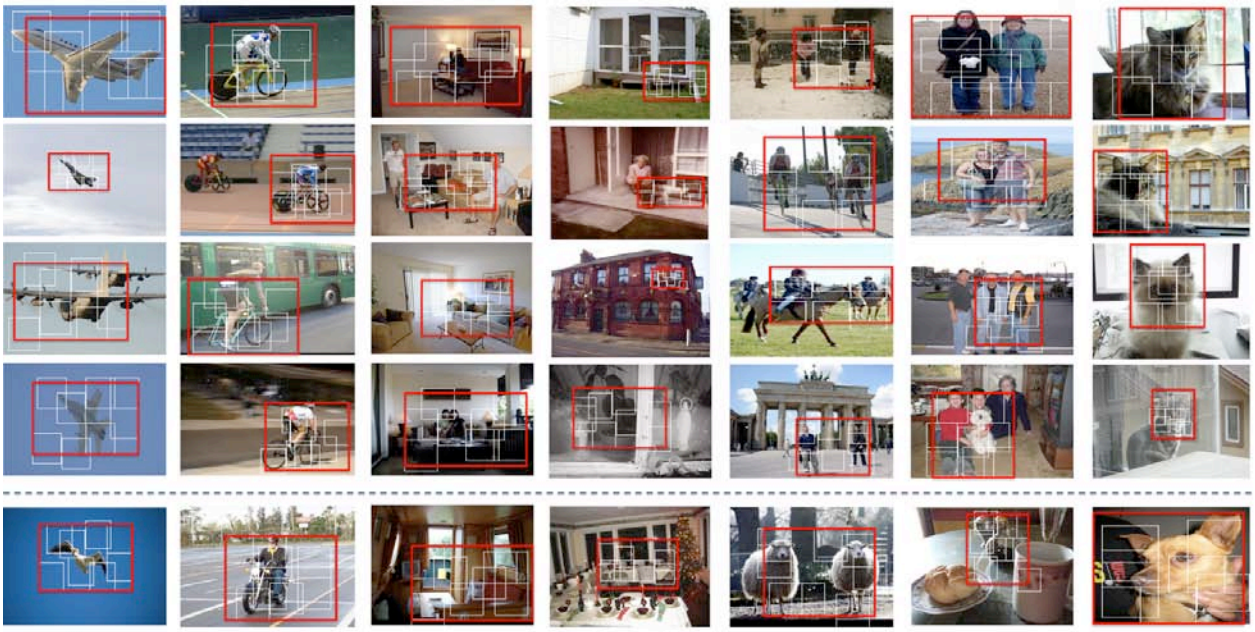


Figure 6. Examples of the detected contextual meta-objects (CMO). Each column shows the CMO detections for a specific category. From left to right: aeroplane, bicycle, sofa, potted-plant, person, person, cat. Each image shows the highest scoring CMO detection. Red box indicates the CMO bounding-box, while the white boxes represent the region-filters within the CMO as learnt by the deformable parts-based model [9]. The first four rows show true-positive detections, and the last row shows false-positive detections. Please refer to the authors' webpages for more results.

tively and automatically tuned for different categories, and capture scene-level, inter-object as well as intra-object interactions. We cast the problem of learning our proposed contextual cue into that of learning object models. This allows us to utilize any off-the-shelf object detector to learn our proposed "contextual meta-objects". We present convincing quantitative and qualitative results on the challenging PASCAL VOC 2007 dataset, where we improve on the performance of several object detectors, and compare favorably to existing sources of context. The benefits of the adaptive granularity at which we extract context, and the potential of our cue to provide complementary information in addition to existing cues are also demonstrated. These improvements do not rely on advanced modeling techniques or learning algorithms, and intelligently leverage existing technology, making them widely accessible.

Acknowledgements: This research was supported in part by the National Science Foundation under IIS-1115719.

References

[1] MSRC 21-class Dataset. http://research.microsoft.com/en-us/projects/objectclassrecognition/.
[2] M. Blaschko and C. Lampert. Object localization with global and local context kernels. In *BMVC*, 2009.
[3] M. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.
[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.
[6] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009.
[7] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge. http://pascallin.ecs.soton.ac.uk/challenges/VOC/.
[8] P. Felzenszwalb, R. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.
[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI, IEEE Transactions on*, 32(9):1627 –1645, sep. 2010.
[10] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *CVPR*, 2010.
[11] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, Anchorage, AK, 2008.
[12] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.
[13] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.
[14] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.
[15] Y. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *CVPR*, 2010.
[16] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vision*, 77:259–289, May 2008.
[17] J. Lim, P. Arbel andez, C. Gu, and J. Malik. Context by region ancestry. In *ICCV*, 2009.
[18] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *NIPS*, 2003.
[19] D. Parikh, C. Zitnick, and T. Chen. From appearance to context-based recognition: Dense labeling in small images. In *CVPR*, 2008.
[20] D. Parikh, C. Zitnick, and T. Chen. Unsupervised learning of hierarchical spatial structures in images. In *CVPR*, 2009.
[21] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010.
[22] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.
[23] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. In *CVPR*, 2009.
[24] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.
[25] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.
[26] A. Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, 2003.
[27] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol Rev*, 113(4), 2006.
[28] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.
[29] L. Wolf and S. Bileschi. A critical view of context. *Int. J. Comput. Vision*, 69:251–261, August 2006.
[30] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.