This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Object-Proposal Evaluation Protocol is 'Gameable'

Neelima Chavali*† Harsh Agrawal* Aroma Mahendru* Dhruv Batra
Virginia Tech
{gneelima, harsh92, maroma, dbatra}@vt.edu

Abstract

Object proposals have quickly become the de-facto pre-processing step in a number of vision pipelines (for object detection, object discovery, and other tasks). Their performance is usually evaluated on partially annotated datasets. In this paper, we argue that the choice of using a partially annotated dataset for evaluation of object proposals is problematic – as we demonstrate via a thought experiment, the evaluation protocol is 'gameable', in the sense that progress under this protocol does not necessarily correspond to a "better" category independent object proposal algorithm.

To alleviate this problem, we: (1) Introduce a nearly-fully annotated version of PASCAL VOC dataset, which serves as a test-bed to check if object proposal techniques are overfitting to a particular list of categories. (2) Perform an exhaustive evaluation of object proposal methods on our introduced nearly-fully annotated PASCAL dataset and perform cross-dataset generalization experiments; and (3) Introduce a diagnostic experiment to detect the bias capacity *in an object proposal algorithm. This tool circumvents the need to collect a densely annotated dataset, which can be expensive and cumbersome to collect. Finally, we have released an easy-to-use toolbox which combines various publicly available implementations of object proposal algorithms which standardizes the proposal generation and evaluation so that new methods can be added and evaluated on different datasets. We hope that the results presented in the paper will motivate the community to test the category independence of various object proposal methods by carefully choosing the evaluation protocol.*

1. Introduction

In the last few years, the Computer Vision community has witnessed the emergence of a new class of techniques called *Object Proposal* algorithms [1–11].

Object proposals are a set of candidate regions or bounding boxes in an image that may potentially contain an object.

Object proposal algorithms have quickly become the de-facto pre-processing step in a number of vision pipelines – object detection [12–21], segmentation [22–26], object discovery [27–30], weakly supervised learning of object-object interactions [31, 32], content aware media retargeting [33], action recognition in still images [34] and visual tracking [35, 36]. Of all these tasks, object proposals have been particularly successful in object detection systems. For example, *nearly all top-performing entries* [13, 37–39] in the ImageNet Detection Challenge 2014 [40] used object proposals. They are preferred over the formerly used sliding window paradigm due to their computational efficiency. Objects present in an image may vary in location, size, and aspect ratio. Performing an exhaustive search over such a high dimensional space is difficult. By using object proposals, computational effort can be focused on a small number of candidate windows.

The focus of this paper is the protocol used for evaluating object proposals. Let us begin by asking – *what is the purpose of an object proposal algorithm?*

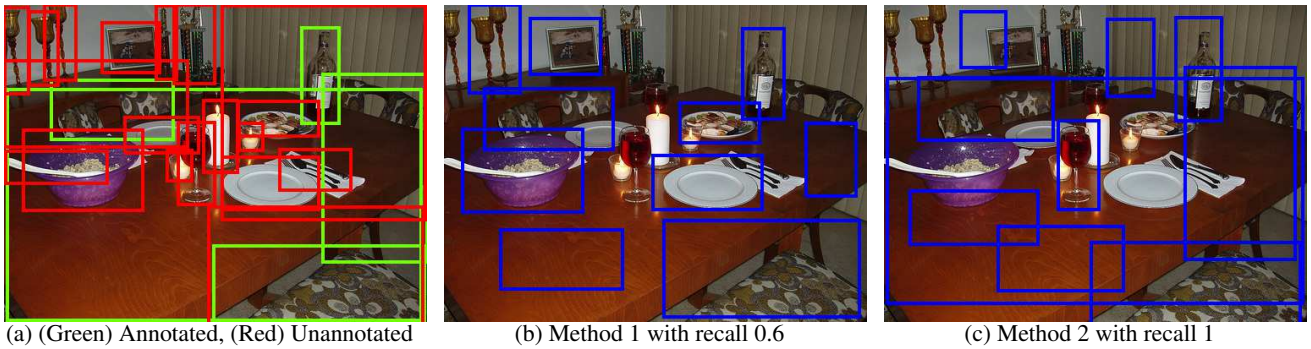
In early works [2, 4, 6], the emphasis was on *category independent object proposals*, where the goal is to identify instances of *all* objects in the image irrespective of their category. While it can be tricky to precisely define what an "object" is[1], these early works presented cross-category evaluations to establish and measure category independence.

More recently, object proposals are increasingly viewed as *detection proposals* [1, 8, 11, 42] where the goal is to improve the object detection pipeline, focusing on a chosen set of object classes (*e.g.* ~20 PASCAL categories). In fact, many modern proposal methods are learning-based [9–11, 42–46] where the definition of an "object" is the set of annotated classes in the dataset. This increasingly blurs the boundary between a proposal algorithm and a detector.

Notice that the former definition has an emphasis on object discovery [27, 28, 30], while the latter definition emphasises on the ultimate performance of a detection pipeline. Surprisingly, despite the two different goals of 'object pro-

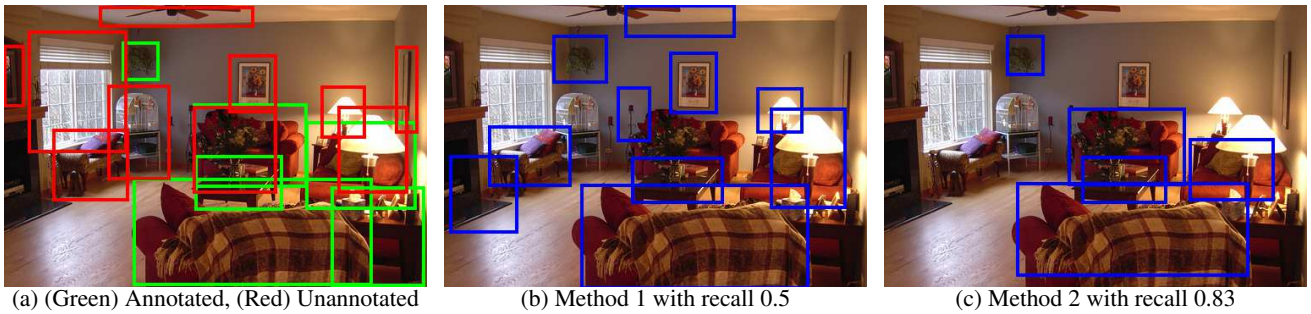
*Equal contribution.
†Now at Amgen Inc.

[1]Most category independent object proposal methods define an object as "stand-alone thing with a well-defined closed-boundary". For "thing" *vs*. "stuff" discussion, see [41].



(a) (Green) Annotated, (Red) Unannotated (b) Method 1 with recall 0.6 (c) Method 2 with recall 1

Figure 1: (a) shows PASCAL annotations natively present in the dataset in green. Other objects that are not annotated but present in the image are shown in red; (b) shows Method 1 and (c) shows Method 2. Method 1 visually seems to recall more categories such as plates, glasses, *etc*. that Method 2 missed. Despite that, the computed recall for Method 2 is higher because it recalled all instances of PASCAL categories that were present in the ground truth. Note that the number of proposals generated by both methods is equal in this figure.



(a) (Green) Annotated, (Red) Unannotated (b) Method 1 with recall 0.5 (c) Method 2 with recall 0.83

Figure 2: (a) shows PASCAL annotations natively present in the dataset in green. Other objects that are not annotated but present in the image are shown in red; (b) shows Method 1 and (c) shows Method 2. Method 1 visually seems to recall more categories such as lamps, picture, *etc*. that Method 2 missed. Clearly the recall for Method 1 *should* be higher. However, the calculated recall for Method 2 is significantly higher, which is counter-intuitive. This is because Method 2 recalls more PASCAL category objects.

posal,' there exists only a single evaluation protocol:

1. Generate proposals on a dataset: The most commonly used dataset for evaluation today is the PASCAL VOC [47] detection set. Note that this is a *partially annotated* dataset where only the 20 PASCAL category instances are annotated.
2. Measure the performance of the generated proposals: typically in terms of 'recall' of the annotated instances. Commonly used metrics are described in Section 3.

The central thesis of this paper is that the current evaluation protocol for object proposal methods is suitable for object detection pipeline but is a *'gameable' and misleading protocol* for category independent tasks. By evaluating only on a specific set of object categories, we fail to capture the performance of the proposal algorithms on *all the remaining object categories that are present in the test set, but not annotated in the ground truth*.

Figs. 1, 2 illustrate this idea on images from PASCAL VOC 2010. Column (a) shows the ground-truth object annotations (in green, the annotations natively present in the dataset for the 20 PASCAL categories –'chairs', 'tables', 'bottles', *etc*.; in red, the annotations that we added to the dataset by marking object such as 'ceiling fan', 'table lamp', 'window', *etc*. originally annotated 'background' in the dataset). Columns (b) and (c) show the outputs of two object proposal methods. Top row shows the case when both methods produce the same number of proposals; bottom row shows unequal number of proposals. We can see that proposal method in Column (b) seems to be more "complete", in the sense that it recalls or discovers a large number of instances. For instance, in the top row it detects a number of non-PASCAL categories ('plate', 'bowl', 'picture frame', *etc*.) but misses out on finding the PASCAL category 'table'. In both rows, the method in Column (c) is reported as achieving a higher recall, *even in the bottom row, when it recalls strictly fewer objects, not just different ones*. The reason is that Column (c) recalls/discovers instances of the 20 PASCAL categories, which are the only ones annotated in the dataset. Thus, Method 2 appears to be a *better* object proposal generator simply because it focuses on the annotated categories in the dataset.

While intuitive (and somewhat obvious) in hindsight, we believe this is a crucial finding because it makes the current protocol *'gameable'* or susceptible to manipulation (both intentional and unintentional) and misleading for measuring improvement in category independent object proposals.

Some might argue that if the end task is to detect a cer-

tain set of categories (20 PASCAL or 80 COCO categories) then it is enough to evaluate on them and there is no need to care about other categories which are not annotated in the dataset. We agree, but it is important to keep in mind that object detection is not the only application of object proposals. There are other tasks for which it is important for proposal methods to generate category independent proposals. For example, in semi/unsupervised object localization [27–30] the goal is to identify all the objects in a given image that contains many object classes without any specific target classes. In this problem, there are no image-level annotations, an assumption of a single dominant class, or even a known number of object classes [28]. Thus, in such a setting, using a proposal method that has tuned itself to 20 PASCAL objects would not be ideal – in the worst case, we may not discover any new objects. As mentioned earlier, there are many such scenarios including learning object-object interactions [31, 32], content aware media retargeting [33], visual tracking [36], *etc*.

To summarize, the contributions of this paper are:

- We report the 'gameability' of the current object proposal evaluation protocol.
- We demonstrate this 'gameability' via a simple thought experiment where we propose a 'fraudulent' object proposal method that *significantly outperforms all existing object proposal techniques* on current metrics, but would under any no circumstances be considered a category independent proposal technique. As a side contribution of our work, we present a simple technique for producing state-of-art object proposals.
- After establishing the problem, we propose three ways of improving the current evaluation protocol to measure the category independence of object proposals:
  1. evaluation on *fully* annotated datasets,
  2. cross-dataset evaluation on *densely* annotated datasets.
  3. a new evaluation metric that quantifies the *bias capacity* of proposal generators.

  For the first test, we introduce a nearly-fully annotated PASCAL VOC 2010 where we annotated *all instances of all object categories* occurring in the images.
- We thoroughly evaluate existing proposal methods on this nearly-fully and two densely annotated datasets.
- We have released release all code and data for experiments[2], and an object proposals library that allows for comparison of popular object proposal techniques.

2. Related Work

Types of Object Proposals: Object proposals can be broadly categorized into two categories:

- **Window scoring**: In these methods, the space of all possible windows in an image is sampled to get a subset of the windows (*e.g.*, via sliding window). These windows are then scored for the presence of an object based on the image features from the windows. The algorithms that fall under this category are [1, 4, 5, 10, 45, 48].
- **Segment based**: These algorithms involve over-segmenting an image and merging the segments using some strategy. These methods include [2, 3, 6–9, 11, 44, 46, 49]. The generated region proposals can be converted to bounding boxes if needed.

Beyond RGB proposals: Beyond the ones listed above, a wide variety of algorithms fall under the umbrella of 'object proposals'. For instance, [50–54] used spatio-temporal object proposals for action recognition, segmentation and tracking in videos. Another direction of work [55–57] explores use of RGB-D cuboid proposals in an object detection and semantic segmentation in RGB-D images. While the scope of this paper is limited to proposals in RGB images, the central thesis of the paper (*i.e.*, gameability of the evaluation protocol) is broadly applicable to other settings.

Evaluating Proposals: There has been a relatively limited analysis and evaluation of proposal methods or the proposal evaluation protocol. Hosang *et al*. [58] focus on evaluation of object proposal algorithms, in particular the stability of such algorithms on parameter changes and image perturbations. Their works shows that a large number of category independent proposal algorithms indeed generalize well to non-PASCAL categories, for instance in the ImageNet 200 category detection dataset [40]. Although these findings are important (and consistent with our experiments), they are unrelated to the 'gameability' of the evaluation protocol. In [59], authors present an analysis of various proposal methods regarding proposal repeatability, ground truth annotation recall, and their impact on detection performance. They also introduced a new evaluation metric (Average Recall). Their argument for a new metric is the need for a better localization between generated proposals and ground truth. While this is a valid and significant concern, it is orthogonal to the'gameability' of the evaluation protocol, which to the best of our knowledge has not been previously addressed. Another recent related work is [60], which analyzes various methods in segment-based object proposals, focusing on the challenges faced when going from PASCAL VOC to MS COCO. They also analyze how aligned the proposal methods are with the bias observed in MS COCO towards small objects and the center of the image and propose a method to boost their performance. Although there is a discussion about biases in datasets but it is unlike our theme, which is 'gameability' due to these biases. As stated earlier, while early papers [2,4,6] reported cross-dataset or cross-category generalization experiments similar to ones reported in this paper, with the trend of learning-based proposal methods, these experiments and concerns seem to have fallen out of standard practice, which we show is problematic.

[2]Data and code can be accessed at: https://filebox.ece.vt.edu/~aroma/web/object-proposals.html

3. Evaluating Object Proposals

Before we describe our evaluation and analysis, let us first look at the object proposal evaluation protocol that is widely used today. The following two factors are involved:

1. **Evaluation Metric**: The metrics used for evaluating object proposals are all typically functions of intersection over union (IOU) (or Jaccard Index) between generated proposals and ground-truth annotations. For two boxes/regions b_i and b_j, IOU is defined as:

$$\text{IOU}(b_i, b_j) = \frac{area(b_i \cap b_j)}{area(b_i \cup b_j)} \quad (1)$$

The following metrics are commonly used:

• **Recall @ IOU Threshold** t: For each ground-truth instance, this metric checks whether the 'best' proposal from list L has IOU greater than a threshold t. If so, this ground truth instance is considered 'detected' or 'recalled'. Then average recall is measured over all the ground truth instances:

$$\text{Recall} \,@\, t = \frac{1}{|G|} \sum_{g_i \in G} \text{I}\left[\max_{l_j \in L} \text{IOU}(g_i, l_j) > t\right], \quad (2)$$

where $I[\cdot]$ is an indicator function for the logical preposition in the argument. Object proposals are evaluated using this metric in two ways:

– plotting Recall-*vs*.-#proposals by fixing t

– plotting Recall-*vs*.-t by fixing the #proposals in L.

• **Area Under the recall Curve (AUC)**: AUC summarizes the area under the Recall-*vs*.-#proposals plot for different values of t in a single plot. This metric measures AUC-*vs*.-#proposals. It is also plotted by varying #proposals in L and plotting AUC-vs-t.

• **Volume Under Surface (VUS)**: This measures the average recall by linearly varying t and varying the #proposals in L on either linear or log scale. Thus it merges both kinds of AUC plots into one.

• **Average Best Overlap (ABO)**: This metric eliminates the need for a threshold. We first calculate the overlap between each ground truth annotation $g_i \in G$, and the 'best' object hypotheses in L. ABO is calculated as the average:

$$\text{ABO} = \frac{1}{|G|} \sum_{g_i \in G} \max_{l_j \in L} \text{IOU}(g_i, l_j) \quad (3)$$

ABO is typically is calculated on a per class basis. Mean Average Best Overlap (MABO) is defined as the mean ABO over all classes.

• **Average Recall (AR)**: This metric was recently introduced in [59]. Here, average recall (for IOU between 0.5 to 1)-*vs*.-#proposals in L is plotted. AR also summarizes proposal performance across different values of t. AR was shown to correlate with ultimate detection performance better than other metrics.

2. **Dataset**: The most commonly used datasets are the PASCAL VOC [47] detection datasets. Note that these are *partially annotated* datasets where only the 20 PASCAL category instances are annotated. Recently analyses have been shown on ImageNet [58], which has more categories annotated than PASCAL, but is still a partially annotated dataset.

4. A Thought Experiment: How to Game the Evaluation Protocol

Let us conduct a thought experiment to demonstrate that the object proposal evaluation protocol can be 'gamed'.

Imagine yourself reviewing a paper claiming to introduce a new object proposal method – called DMP.

Before we divulge the details of DMP, consider the performance of DMP shown in Fig. 3 on the PASCAL VOC 2010 dataset, under the AUC-*vs*.-#proposals metric.

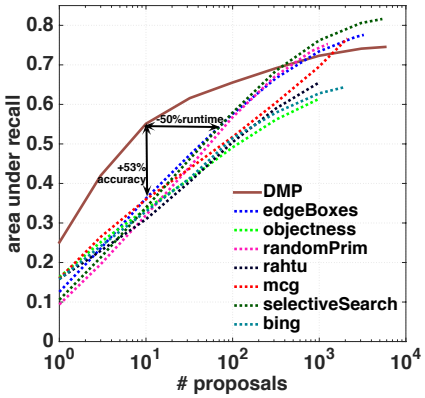


Figure 3: Performance of different object proposal methods (dashed lines) and our proposed 'fraudulent' method (DMP) on the PASCAL VOC 2010 dataset. We can see that DMP *significantly* outperforms all other proposal generators. See text for details.

As we can clearly see, the proposed method DMP *significantly* exceeds all existing proposal methods [1–6, 8, 10, 11] (which seem to have little variation over one another). The improvement at some points in the curve (*e.g.*, at M=10) seems to be *an order of magnitude* larger than all previous incremental improvements reported in the literature! In addition to the gain in AUC at a fixed M, DMPs also achieves the same AUC (0.55) at an *order of magnitude fewer* number of proposals (M=10 *vs*. M= 50 for edgeBoxes [1]). Thus, fewer proposals need to be processed by the ensuing detection system, resulting in an equivalent run-time speedup. This seems to indicate that a significant progress has been made in the field of generating object proposals.

So what is our proposed state-of-art technique DMP?
It is a mixture-of-experts model, consisting of 20 experts, where each expert is a deep feature (fc7)-based [61] objectness detector. At this point, you, the savvy reader, are probably already beginning to guess what we did.

DMP stands for 'Detector Masquerading as Proposal generator'. We trained object detectors for the 20 PASCAL categories (in this case with RCNN [12]), and then used these 20 detectors to produce the top-M most confident detections (after NMS), and declared them to be 'object proposals'.

The point of this experiment is to demonstrate the following fact – clearly, no one would consider a collection of 20 object detectors to be a category independent object proposal method. However, our existing evaluation protocol declared the union of these top-M detections to be state-of-the-art.

Why did this happen? Because the protocol today involves evaluating a proposal generator on *a partially annotated* dataset such as PASCAL. The protocol does not reward recall of non-PASCAL categories; in fact, early recall (near the top of the list of candidates) of non-PASCAL objects results in a penalty for the proposal generator! As a result, a proposal generator that tunes itself to these 20 PASCAL categories (either explicitly via training or implicitly via design choices or hyper-parameters) will be declared a better proposal generator when it may not be (as illustrated by DMP). Notice that as learning-based object proposal methods improve on this metric, "in the limit" *the best object proposal technique is a detector for the annotated categories*, similar to our DMP. Thus, we should be cautious of methods proposing incremental improvements on this protocol – improvements on this protocol do not necessarily lead to a better category independent object proposal method.

This thought experiment exposes the inability of the existing protocol to evaluate category independence.

5. Evaluation on Fully and Densely Annotated Datasets

As described in the previous section, the problem of 'gameability' is occurring due to the evaluation of proposal methods on partially annotated datasets. An intuitive solution would be evaluating on a *fully* annotated dataset.

In the next two subsections, we evaluate the performance of 7 popular object proposal methods [1, 3–6, 8, 10] and two DMPs (RCNN [12] and DPM [63]) on one nearly-fully and two densely annotated datasets containing many more object categories. This is to quantify how much the performance of our 'fraudulent' proposal generators (DMPs) drops once the bias towards the 20 PASCAL categories is diminished (or completely removed).

We begin by *creating* a nearly-fully annotated dataset by building on the effort of PASCAL Context [62] and evaluate on this nearly-fully annotated modified instance level PASCAL Context; followed by cross-dataset evaluation on other partial-but-densely annotated datasets MS COCO [64] and NYU-Depth V2 [65].

Experimental Setup: On MS COCO and PASCAL Context datasets we conducted experiments as follows:

- Use the existing evaluation protocol for evaluation, *i.e.*, evaluate only on the 20 PASCAL categories.
- Evaluate on all the annotated classes.
- For the sake of completeness, we also report results on all the classes except the PASCAL 20 classes.[3]

Training of DMPs: The two DMPs we use are based on two popular object detectors - DPM [63] and RCNN [12]. We train DPM on 20 PASCAL categories and use it as an object proposal method. To generate large number of proposals, we chose a low value of threshold in Non-Maximum Suppression (NMS). Proposals are generated for each category and a score is assigned to them by the corresponding DPM for that category. These proposals are then merge-sorted on the basis of this score. Top M proposals are selected from this sorted list where M is the number of proposals to be generated.

Another (stronger) DMP is RCNN which is a detection pipeline that uses 20 SVMs (each for one PASCAL category) trained on deep features (fc7) [61] extracted on selective search boxes. Since RCNN itself uses selective search proposals, it should be viewed as a trained *reranker* of selective search boxes. As a consequence, it ultimately equals selective search performance once the number of candidates becomes large. We used the pretrained SVM models released with the RCNN code, which were trained on the 20 classes of PASCAL VOC 2007 trainval set. For every test image, we generate the Selective Search proposals using the 'FAST' mode and calculate the 20 SVM scores for each proposal. The 'objectness' score of a proposal is then the maximum of the 20 SVM scores. All the proposals are then sorted by this score and top M proposals are selected.[4]

Object Proposals Library[2]: To ease the process of carrying out the experiments, we created an open source, easy-to-use object proposals library. This can be used to seamlessly generate object proposals using all the existing algorithms [1–9] (for which the Matlab code has been released by the respective authors) and evaluate these proposals on any dataset using the commonly used metrics.

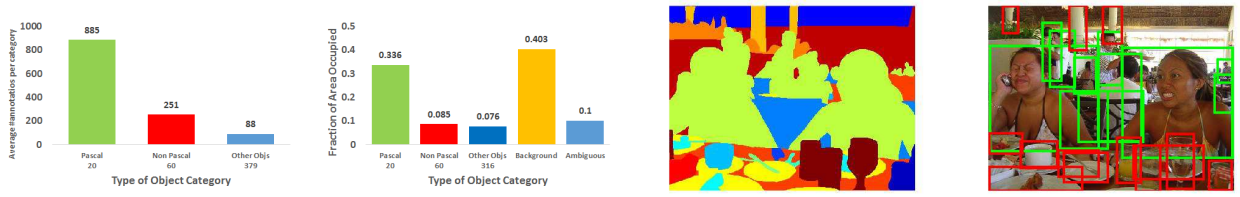
5.1. Fully Annotated Dataset

PASCAL Context: This dataset ()introduced by Mottaghi *et al*. [62]) contains additional annotations for PASCAL VOC 2010 dataset [66]. The annotations are semantic segmentation maps, where *every single pixel* previously annotated 'background' in PASCAL was assigned a category label. In total, annotations have been provided for 459 categories. This includes the original 20 PASCAL categories and new classes such as keyboard, fridge, picture, cabinet.

Unfortunately, the dataset contains only category-level semantic segmentations. For our task, we needed instance-level annotations, which can't be reliably extracted from category-level segmentation masks.

[3] On NYU-Depth V2 evaluation is done on all categories. This is because only 8 PASCAL categories are present in this dataset.

[4] It was observed that merge-sorting calibrated/rescaled SVM scores led to inferior performance as compared to merge-sorting without rescaling.



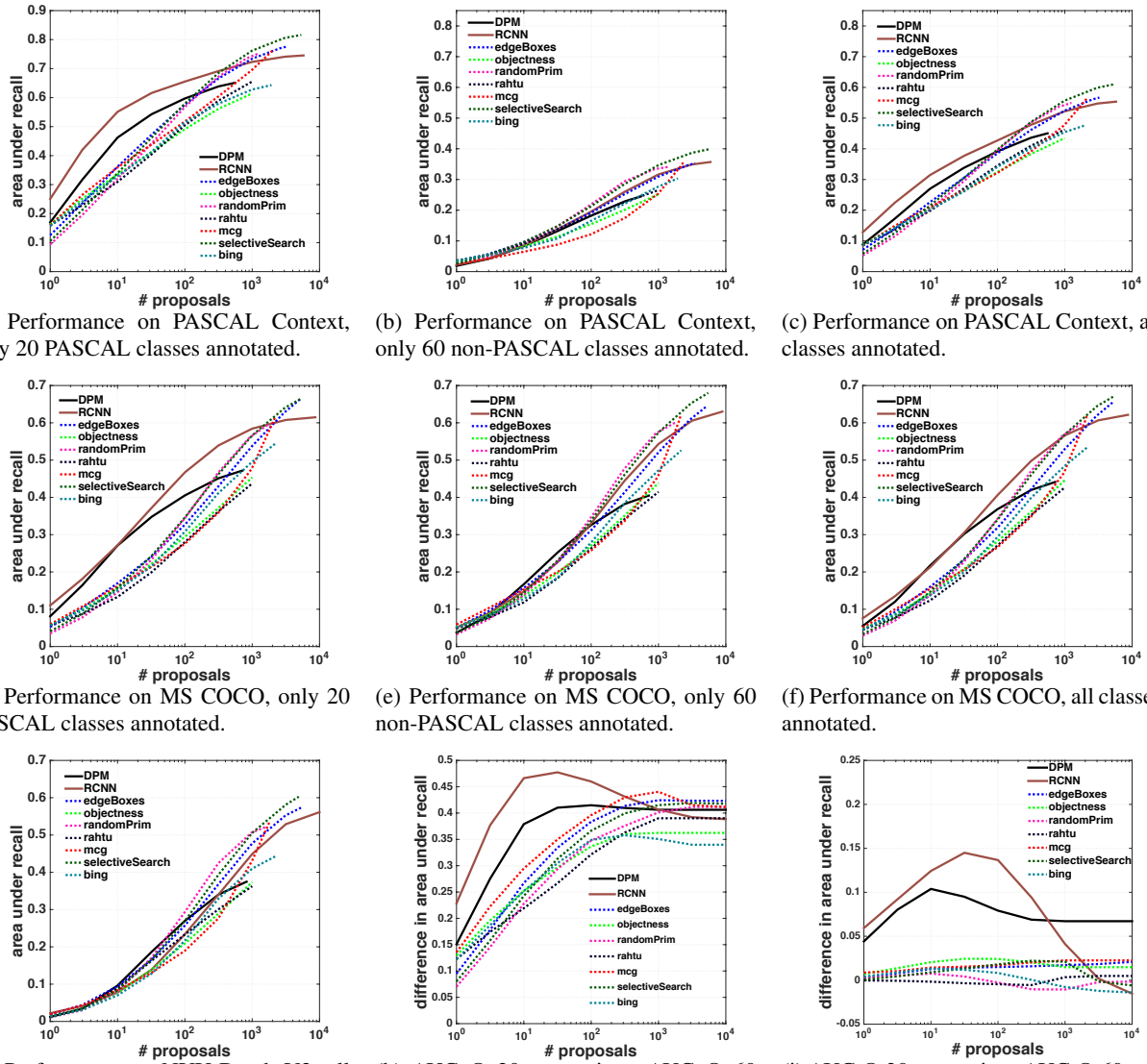
(a) Average #annotations for different categories.

(b) Fraction of image-area covered by different categories.

(c) PASCAL Context annotations [62].

(d) Our augmented annotations.

Figure 4: (a),(b) Distribution of object classes in PASCAL Context with respect to different attributes. (c),(d) Augmenting PASCAL Context with instance-level annotations. (Green = PASCAL 20 categories; Red = new objects)



(a) Performance on PASCAL Context, only 20 PASCAL classes annotated.

(b) Performance on PASCAL Context, only 60 non-PASCAL classes annotated.

(c) Performance on PASCAL Context, all classes annotated.

(d) Performance on MS COCO, only 20 PASCAL classes annotated.

(e) Performance on MS COCO, only 60 non-PASCAL classes annotated.

(f) Performance on MS COCO, all classes annotated.

(g) Performance on NYU-Depth V2, all classes annotated

(h) AUC @ 20 categories - AUC @ 60 categories on PASCAL Context.

(i) AUC @ 20 categories - AUC @ 60 categories on MS COCO.

Figure 5: Performance of different methods on PASCAL Context, MS COCO and NYu Depth-V2 with different sets of annotations.

Creating Instance-Level Annotations for PASCAL Context[2]: Thus, we created instance-level bounding box annotations for all images in PASCAL Context dataset. First, out of the 459 category labels in PASCAL Context, we identified 396 categories to be 'things', and ignored the remaining

'stuff' or 'ambiguous' categories[5] – neither of these lend themselves to bounding-box-based object detection. See supplement[6] for details.

We selected the 60 most frequent non-PASCAL categories from this list of 'things' and manually annotated all their instances. Selecting only top 60 categories is a reasonable choice because the average per category frequency in the dataset for all the other categories (even after including background/ambiguous categories) was roughly one third as that of the chosen 60 categories (Fig. 4a). Moreover, the percentage of pixels in an image left unannotated (as 'background') drops from 58% in original PASCAL to 50% in our nearly-fully annotated PASCAL Context. This manual annotation was performed with the aid of the semantic segmentation maps present in the PASCAL Context annotations. Examples annotations are shown in Fig. 4d. For detailed statistics, see supplement[6].

Results and Observations: We now explore how changes in the dataset and annotated categories affect the results of the thought experiment from Section 4. Figs. 5a, 5b, 5c, 5h compare the performance of DMPs with a number of existing proposal methods [1–6, 8, 10, 11] on PASCAL Context.

We can see in Column (a) that when evaluated on only 20 PASCAL categories DMPs trained on these categories appear to significantly outperform all proposal generators. However, we can see that they are not category independent because they suffer a big drop in performance when evaluated on 60 non-PASCAL categories in Column (b). Notice that on PASCAL context, *all proposal generators* suffer a drop in performance between the 20 PASCAL categories and 60 non-PASCAL categories. We hypothesize that this due to the fact that the non-PASCAL categories tend to be generally smaller than the PASCAL categories (which were the main targets of the dataset curators) and hence difficult to detect. But this could also be due to the reason that authors of these methods made certain choices while designing these approaches which catered better to the 20 annotated categories. However, the key observation here (as shown in Fig. 5h) is that DMPs suffer the biggest drop. This drop is much greater than all the other approaches. It is interesting to note that due to the ratio of instances of 20 PASCAL categories vs other 60 categories, DMPs continue to slightly outperform proposal generators when evaluated on all categories, as shown in Column (c).

5.2. Densely Annotated Datasets

Besides being expensive, "full" annotation of images is somewhat ill-defined due to the hierarchical nature of object semantics (*e.g.* are object-parts such as bicycle-wheel, windows in a building, eyes in a face, *etc.* also objects?). One way to side-step this issue is to use datasets with dense annotations (albeit at the same granularity) and conduct cross-dataset evaluation.

MS COCO: Microsoft Common Objects in Context (MS COCO) dataset [64] contains 91 common object classes (82 of them having more than 5,000 labeled instances). It not only has significantly higher number of instances per class than PASCAL, but also more object instances per image (7.7) as compared to ImageNet (3.0) and PASCAL (2.3).

NYU-Depth V2: NYU-Depth V2 dataset [65] is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras. It features 1449 densely labeled pairs of aligned RGB and depth images with instance-level annotations. We used these 1449 densely annotated RGB images for evaluating object proposal algorithms. To the best of our knowledge, this is the first paper to compare proposal methods on such a dataset.

Results and Observations: Figs. 5d, 5e, 5f, 5i show a plot similar to PASCAL Context on MS COCO. Again, DMPs outperform all other methods on PASCAL categories but fail to do so for the Non-PASCAL categories. Fig. 5g shows results for NYU-Depth V2. See that when many classes in the test dataset are not PASCAL classes, DMPs tend to perform poorly, although it is interesting that the performance is still not as poor as the worst proposal generators. Results on other evaluation criteria are in the supplement[6].

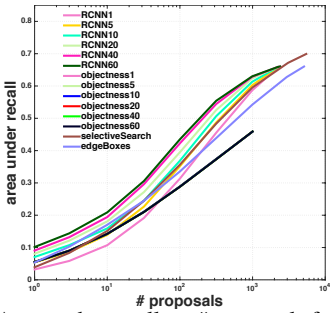
6. Bias Inspection

So far, we have discussed two ways of detecting 'gamebility' – evaluation on nearly-fully annotated dataset and cross-dataset evaluation on densely annotated datasets. Although these methods are fairly useful for bias detection, they have certain limitations. Datasets can be unbalanced. Some categories can be more frequent than others while others can be hard to detect (due to choices made in dataset collection). These issues need to be resolved for perfectly unbiased evaluation. However, generating unbiased datasets is an expensive and time-consuming process. Hence, to detect the bias without getting unbiased datasets, we need a method which can measure performance of proposal methods in a way that category specific biases can be accounted for and the extent or the *capacity* of this bias can be measured. We introduce such a method in this section.

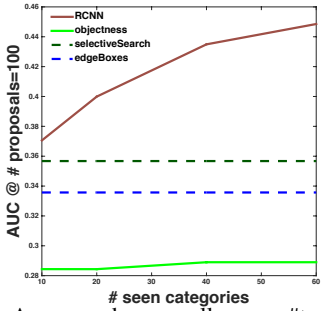
6.1. Assessing Bias Capacity

Many proposal methods [9–11, 42–46] rely on explicit training to learn an "objectness" model, similar to DMPs. Depending upon which, how many categories they are trained on, these methods could have a biased view of "objectness". One way of measuring the *bias capacity* in a proposal method to plot the performance *vs.* the number of 'seen' categories while evaluating on some held-out set. A method that involves little or no training will be a flat curve on this plot. Biased methods such as DMPs will get better and better as more categories are seen in training. Thus, this analysis can help us find biased or 'gamebility-prone'

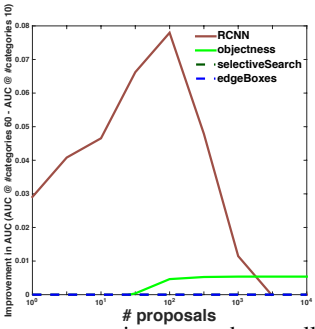
[5] *e.g.*, a 'tree' may be a 'thing' or 'stuff' subject to camera viewpoint.
[6] See Appendix in: http://arxiv.org/abs/1505.05836



(a) Area under recall *vs*. # proposals for various #seen categories



(b) Area under recall *vs*. #training-categories.



(c) Improvement in area under recall from #seen categories =10 to 60 *vs*. # proposals.

Figure 6: Performance of RCNN and other proposal generators vs number of object categories used for training. We can see that RCNN has the most 'bias capacity' while the performance of other methods is nearly (or absolutely) constant.

methods like DMPs that are/can be tuned to specific classes. To the best of our knowledge, no previous work has attempted to measure bias capacity by varying the number of 'object' categories seen at training time. In this experiment, we compared the performance of one DMP method (RCNN), one learning-based proposal method (Objectness), and two non learning-based proposal methods (Selective Search [8], EdgeBoxes [1]) as a function of the number of 'seen' categories (the categories trained on[7]) on MS COCO [64] dataset. Method names 'RCNNTrainN', 'objectnessTrainN' indicate that they were trained on images that contain annotations for only N categories (50 instances per category). Total number of images for all 60 categories was ~2400 (because some images contain >1 object). Once trained, these methods were evaluated on a randomly-chosen set of ~500 images, which had annotations for all 60 categories.

Fig. 6a shows Area under Recall *vs*. #proposals curve for learning-based methods trained on different sets of categories. Fig. 6b and Fig. 6c show the variation of AUC *vs*. # seen categories and improvement due to increase in training categories (from 10 to 60) *vs*. #proposals respectively, for RCNN and objectness when trained on different sets of categories. The key observation to make here is that with even a modest increase in 'seen' categories with the same amount of increased training data, performance improvement of RCNN is significantly more than objectness. Selective Search [8] and edgeBoxes [1] are the dashed straight lines since there is no training involved.

These results clearly indicate that as RCNN sees more categories, its performance improves. One might argue that the reason might be that the method is learning more 'objectness' as it is seeing more data. However, as discussed above, the increase in the dataset size is marginal (~40 images per category) and hence it unlikely that such a significant improvement is observed due to that. Thus, it is reasonable to conclude that this improvement is because the method is learning class specific features.

[7]The seen categories are picked in the order they are listed in MS COCO dataset (*i.e*., no specific criterion was used).

Thus, this approach can be used to reason about 'gameability-prone' and 'gameability-immune' proposal methods without creating an expensive fully annotated dataset. We believe this simple but effective diagnostic experiment would help to detect and thus contribute in managing the category specific bias in all learning-based methods.

7. Conclusion

To conclude, the main message of this paper is simply this – the current evaluation protocol for object proposal algorithms is not suitable if we view them as category independent object proposal methods (meant to *discover* [27, 28], instances of all categories). By evaluating the 'recall' of instances on a partially annotated dataset, we fail to capture the performance of the proposal algorithm on all the remaining object categories that are present in the test set, but not annotated in the ground truth.

We demonstrate this 'gameability' via a simple thought experiment where we propose a 'fraudulent' object proposal method that outperforms all existing object proposal techniques on current metrics. We introduce a nearly-fully annotated version of PASCAL VOC 2010 where we annotated all instances of 60 object categories other than 20 PASCAL categories occurring in all images. We perform an exhaustive evaluation of object proposal methods on our introduced modified instance level PASCAL dataset and perform cross-dataset generalization experiments on MS COCO and NYU-Depth V2. We have also released an easy-to-use library to evaluate and compare various proposal methods which we think will also be useful. Furthermore, since densely annotating the dataset is a tedious and costly task, we proposed a diagnostic experiment to detect and quantify the bias capacity in object proposal methods.

As modern proposal methods become more learning-based and trained in an end-to-end fashion, it is clear that the distinction between detectors and proposal generators is becoming blurred. With that in mind, it is important to recognize and safeguard against the flaws in the protocol, lest we over-fit as a community to a specific set of object classes.

References

[1] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014. 1, 3, 4, 5, 7, 8

[2] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *PAMI*, vol. 36, no. 2, pp. 222–234, 2014. 1, 3, 4, 5, 7

[3] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014. 1, 3, 4, 5, 7

[4] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *PAMI*, 2012. 1, 3, 4, 5, 7

[5] E. Rahtu, J. Kannala, and M. B. Blaschko, "Learning a category independent object detection cascade," in *ICCV*, 2011. 1, 3, 4, 5, 7

[6] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized prim's algorithm," in *ICCV*, 2013. 1, 3, 4, 5, 7

[7] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *CVPR*, 2014. 1, 3, 5

[8] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, 2013. 1, 3, 4, 5, 7, 8

[9] A. Humayun, F. Li, and J. M. Rehg, "Rigor- recycling inference in graph cuts for generating object regions," in *CVPR*, 2014. 1, 3, 5, 7

[10] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014. 1, 3, 4, 5, 7

[11] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *ECCV*, 2014. 1, 3, 4, 7

[12] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. 1, 5

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, 2014. 1

[14] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov, "Scalable, high-quality object detection," *CoRR*, vol. abs/1412.1441, 2014. 1

[15] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *ICCV*, 2013. 1

[16] R. G. Cinbis, J. Verbeek, and C. Schmid, "Segmentation Driven Object Detection with Fisher Vectors," in *ICCV*, 2013. 1

[17] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," *CoRR*, vol. abs/1312.2249, 2013. 1, 8

[18] A. Kuznetsova, S. Ju Hwang, B. Rosenhahn, and L. Sigal, "Expanding object detector's horizon: Incremental learning framework for object detection in videos," in *CVPR*, 2015. 1

[19] Y.-H. Tsai, O. C. Hamsici, and M.-H. Yang, "Adaptive region pooling for object detection," in *CVPR*, 2015. 1

[20] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, "segdeepm: Exploiting segmentation and context in deep neural networks for object detection," in *CVPR*, 2015. 1

[21] S. He and R. W. Lau, "Oriented object proposals," in *ICCV*, 2015. 1

[22] J. Carreira and C. Sminchisescu, "CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts," *PAMI*, vol. 34, 2012. 1

[23] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. D. Bourdev, and J. Malik, "Semantic segmentation using regions and parts.," in *CVPR*, 2012. 1

[24] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *ECCV*, 2012. 1

[25] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation.," in *CVPR*, 2015. 1

[26] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia, "Semantic object segmentation via detection in weakly labeled video," in *CVPR*, 2015. 1

[27] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *IJCV*, no. 3, pp. 275–293, 2012. 1, 3, 8

[28] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," *CoRR*, vol. abs/1501.06170, 2015. 1, 3, 8

[29] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler, "Active learning and discovery of object categories in the presence of unnameable instances," in *CVPR*, 2015. 1, 3

[30] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images.," in *CVPR*, IEEE, 2013. 1, 3

[31] R. G. Cinbis and S. Sclaroff, "Contextual object detection using set-based classification," in *ECCV*, 2012. 1, 3

[32] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *PAMI*, vol. 34, no. 3, pp. 601–614, 2012. 1, 3

[33] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *ICCV*, 2011. 1, 3

[34] F. Sener, C. Bas, and N. Ikizler-Cinbis, "On recognizing actions in still images via multiple features," in *ECCV*, 2012. 1

[35] N. Wang, S. Li, A. Gupta, and D. Yeung, "Transferring rich feature hierarchies for robust visual tracking," *CoRR*, vol. abs/1501.04587, 2015. 1

[36] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid, "Unsupervised object discovery and tracking in video collections," *CoRR*, vol. abs/1505.03825, 2015. 1, 3

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014. 1

[38] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013. 1

[39] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, *et al.*, "Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection," *arXiv preprint arXiv:1409.3505*, 2014. 1

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," 2014. 1, 3

[41] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. 10th European Conference on Computer Vision*, 2008. 1

[42] P. Krähenbühl and V. Koltun, "Learning to propose objects," in *CVPR*, 2015. 1, 7, 8

[43] H. Kang, M. Hebert, A. A. Efros, and T. Kanade, "Data-driven objectness," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 189–195, 2015. 1, 7

[44] W. Kuo, B. Hariharan, and J. Malik, "Deepbox: Learning objectness with convolutional networks," 2015. 1, 3, 7

[45] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *ICCV*, 2015. 1, 3, 7

[46] P. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," *CoRR*, vol. abs/1506.06204, 2015. 1, 3, 7, 8

[47] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2, 4

[48] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *CVPR*, 2015. 3

[49] C. Wang, L. Zhao, S. Liang, L. Zhang, J. Jia, and Y. Wei, "Object proposal by multi-branch hierarchical segmentation," in *CVPR*, 2015. 3

[50] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, "Spatio-temporal object detection proposals," in *ECCV 2014*, 2014. 3

[51] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Spatio-temporal moving object proposals," *arXiv preprint arXiv:1412.6504*, 2014. 3

[52] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *CVPR*, 2015. 3

[53] I. Misra, A. Shrivastava, and M. Hebert, "Watch and learn: Semi-supervised learning for object detectors from video," in *CVPR*, 2015. 3

[54] Z. Wu, F. Li, R. Sukthankar, and J. M. Rehg, "Robust video segment proposals with painless occlusion handling," in *CVPR*, 2015. 3

[55] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," *arXiv preprint arXiv:1407.5736*, 2014. 3

[56] D. Banica and C. Sminchisescu, "CPMC-3D-O2P: semantic segmentation of RGB-D images using CPMC and second order pooling," *CoRR*, vol. abs/1312.7715, 2013. 3

[57] D. Banica and C. Sminchisescu, "Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images," in *CVPR*, 2015. 3

[58] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?," in *BMVC*, 2014. 3, 4

[59] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?," *arXiv preprint arXiv:1502.05082*, 2015. 3, 4

[60] J. Pont-Tuset and L. Van Gool, "Boosting object proposals: From pascal to coco," in *International Conference on Computer Vision*, 2015. 3

[61] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013. 4, 5

[62] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014. 5, 6

[63] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010. 5

[64] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014. 5, 7, 8

[65] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012. 5, 7

[66] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results." http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 5

[67] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. 8