Detecting avocados to zucchinis: what have we done, and where are we going?

Olga Russakovsky[1], Jia Deng[1], Zhiheng Huang[1], Alexander C. Berg[2], Li Fei-Fei[1]
Stanford University[1] , UNC Chapel Hill[2]

Abstract

The growth of detection datasets and the multiple directions of object detection research provide both an unprecedented need and a great opportunity for a thorough evaluation of the current state of the field of categorical object detection. In this paper we strive to answer two key questions. First, where are we currently as a field: what have we done right, what still needs to be improved? Second, where should we be going in designing the next generation of object detectors?

Inspired by the recent work of Hoiem et al. [10] on the standard PASCAL VOC detection dataset, we perform a large-scale study on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data. First, we quantitatively demonstrate that this dataset provides many of the same detection challenges as the PASCAL VOC. Due to its scale of 1000 object categories, ILSVRC also provides an excellent testbed for understanding the performance of detectors as a function of several key properties of the object classes. We conduct a series of analyses looking at how different detection methods perform on a number of image-level and object-class-level properties such as texture, color, deformation, and clutter. We learn important lessons of the current object detection methods and propose a number of insights for designing the next generation object detectors.

1. Introduction

The field of large-scale categorical object detection is rapidly growing [7, 3, 12], both by developing more robust object representations and also by collecting richer datasets for training and evaluation. In this paper, we ask two questions. Where are we now as a field: what challenges of large-scale object detection have we successfully addressed, and which ones still remain? And where should we be going in building the next generation of object detectors?

In order to begin answering these questions, we first consider the datasets that are used to train and evaluate current recognition and detection systems. There has been rapid progress on this front in computer vision, starting from the perhaps overly simplistic Caltech 101 [8] and moving towards progressively larger scale and more representative of everyday scenes and objects through many datasets, e.g., PASCAL [7], LabelMe [15], TinyImages [19], SUN [24], and ImageNet [4]. The large variety of datasets has sparked a recent trend of performing meta-analyses, such as the work by Torralba and Efros [18] examining the generalizability of existing datasets. They confirm that one of the more varied and generalizable datasets is the PASCAL Visual Object Classes (VOC) challenge [7]. PASCAL VOC has carried the torch of benchmarking advances in object detection for the past 7 years. The dataset and the associated detection challenge inspired and advertised multiple breakthroughs in generic object detection [9, 21, 20].

Hoiem et al. [10] designed and performed a thorough evaluation of several state-of-the-art object detectors on the PASCAL dataset, providing much more detail than single average precision score for each category. This study highlighted some insights into where detection algorithms make mistake – e.g., false positive detections surprisingly rarely occur on the background clutter and much more often appear on objects of similar categories. This kind of analysis offers a way to examine the current state of the field of object detection, but more importantly sheds light on what should be done for designing the next generation of object detectors. While other insights can be obtained from further analyzing the PASCAL dataset, one limitation is that it contains only 20 basic object categories. Therefore, it is harder to do meta-analysis or measure the impact of object properties such as color, texture, real-world size, on the performance of object detectors. Such analysis is important in understanding when detection approaches can be expected to work or to fail, and where more research is needed.

An alternative is to use the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [3] data with 1,000 object classes for benchmarking and analyzing detection. This was not possible until ILSVRC added the localization task in year 2011. There is some concern that the ImageNet data used in ILSVRC is too easy and has a different bias (e.g., more centered objects, less cluttered background) than the PASCAL data. We carefully analyze the ILSVRC data

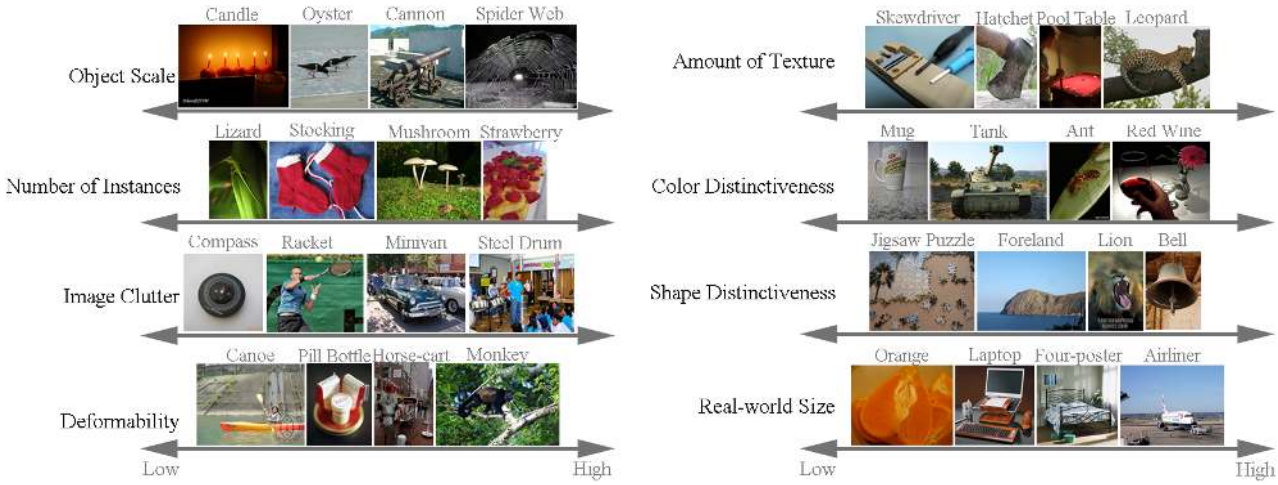

Figure 1. The diversity of ILSVRC along eight dimensions. Please refer to Section 2.2 for definitions of object scale, number of instances and image clutter, and to Section 3.3 for the rest. For each dimension, we show example object categories along the range of that property.

and show that subsets of it can be selected such that clutter, object scale, and other statistics contributing to detection difficulty are well matched to PASCAL while still keeping the benefits of much of the variety and scale of the original ILSVRC dataset.

The large number and variety of object classes in ILSVRC allows us to analyze statistical properties of objects (Figure 1) and their impact on detection algorithms on an unprecedented scale. By utilizing two leading object detection algorithms from the 2012 challenge, we learn several important lessons for future object detection.

- The current recognition systems are very strong at classifying basic-level object categories when trained on large-scale data; however, fine-grained recognition remains very challenging and holds potential for great improvements of detector accuracy (Section 3.1).
- The amount of clutter in images (measured using a novel metric introduced in Section 2.2) is a strong predictor of detection accuracy on a large scale (Section 3.2). While the fact that clutter makes object detection more challenging is intuitive, this finding is particularly interesting considering the recent observation of [10] on PASCAL VOC that false positive detections surprisingly rarely occur on background clutter.
- The object detection system based on the deformable parts-based model (DPM) of [9] is much more robust to factors such as number of object instances per image and object scale than a neural network-based method [12]. However, the neural network system outperforms the DPM in classification and localization accuracy even on the most challenging subsets of ILSVRC modeled after PASCAL data (Section 3.2).
- Man-made objects are significantly more challenging to detect than natural objects, and objects with highly distinctive textures are significantly easier than untextured objects. Importantly, this holds true on a scale of 1,000 classes even when controlling for factors such as object scale and level of clutter. (Section 3.3).

Building upon these key insights we offer design recommendations for future object detector in Section 4.[1]

Related work. Several works have analyzed the effects of factors such as occlusion, variations in aspect ratios and changes of viewpoints, for both specific category detection and general object detection on a small set of categories [23, 22, 11, 6, 5]. Others have provided insight into dataset design [14, 18, 7]. [25] analyzed the relative impact of adding more training data versus building better detection models. The most in-depth analysis of generic object detection to date has been performed on the PASCAL challenge in [7] and especially in [10]. In this paper we go beyond the scope of previous analysis by capturing object statistics which are near impossible to get on smaller scale data.

Notation. Throughout the paper, we use the abbreviation **ILSVRC** to refer specifically to the 2012 ImageNet Large Scale Visual Recognition Challenge. We use **PASCAL** to refer to the 2012 PASCAL Visual Object Classes challenge.

2. Evaluation protocol

We set up the three key components of our analysis of the current state of large-scale object detection. Section 2.1 describes the ILSVRC dataset and the evaluation criteria adapted to the large-scale localization setting [3]. Section 2.2 introduces three quantitative measures of localization difficulty. These are used to compare the statistics of the ILSVRC data to the standard PASCAL benchmark and

[1] More information is available in the supplementary material and at www.image-net.org/challenges/LSVRC/2012/analysis/

Figure 2. The ILSVRC2012 dataset contains many more fine-grained classes compared to the standard PASCAL VOC benchmark; for example, instead of the PASCAL "dog" category there are 120 different breeds of dogs in ILSVRC2012.

then later in the analysis throughout the paper. Section 2.3 describes the detection algorithms chosen as representatives of the current state-of-the-art in large-scale detection.

2.1. ILSVRC classification+localization dataset

Data collection. ImageNet is an image dataset organized according to the WordNet hierarchy [4]. ImageNet provides quality-controlled, human-annotated images illustrating more than 21 thousand concepts. The image collection protocol is similar to that of the PASCAL dataset. Flickr, which is the source of PASCAL data, is also the source of 44% of the ImageNet images. In addition, ImageNet images also come from a variety of other sources, such as Google, Bing and Picsearch image search engines queried in multiple languages. Figure 1 visualizes some of the diversity of ImageNet along several dimensions.

A subset of the ImageNet data is used for ILSVRC, spanning 1000 classes containing both internal nodes and leaf nodes of ImageNet. There is no overlap between the nodes, e.g., since dog breed "fox terrier" is present as one of the categories, the general "dog" category is not. Hoiem et al. [10] showed that many detection errors stem from confusion between objects of similar classes; the ILSVRC dataset provides an unprecedented challenge for detectors in this regard by having many fine-grained classes, e.g., three schnauzer breeds in Figure 2. There are 1.2 million images for training, 50K images for validation, and 100K new images for testing. There are 620K bounding box annotations on the training images (covering about 42% of the data), and an additional 230K for the validation and test images.

Evaluation criteria. The commonly accepted measure of detection accuracy of an algorithm consists of two requirements [7]: (1) all instances of an object class should be correctly localized, where an object instance with a bounding box $\hat{B}$ is considered correctly localized if a window B returned by the algorithm satisfies the intersection over union (IOU) measure: $\text{IOU}(\hat{B}, B) = \frac{\text{area}(\hat{B} \cap B)}{\text{area}(\hat{B} \cup B)} \geq 0.5$, and (2) there are few false positive detections.

The scale of the PASCAL dataset (20 object categories and tens of thousands of images) allows for thorough ground truth annotation of the presence of all the object categories on all images. However, on the ILSVRC scale of 1000 categories and more than a million of images it is infeasible to label every instance of every object in every image. Therefore, a modified measure of localization accuracy is used. Each object class C has a set of images associated with it, and each image is human annotated with bounding boxes $B_1, B_2, \ldots$ indicating the location of *all* instances of this object class. Since additional unannotated object classes may be present, the algorithm is allowed to produce up to 5 annotations per image without incurring a cost for false positive detections.[2] The object class is considered correctly detected if for some proposed annotation (c_i, b_i) with c_i the class label and b_i the bounding box, $c_i = C$ and b_i correctly localizes one of the objects $B_1, B_2, \ldots$ according to the standard IOU measure.

This criteria allows for evaluation of the methods in a large-scale setting where complete human annotation is infeasible, and is similar to evaluating detection recall for a fixed rate of false positives per image which is standard in e.g, pedestrian detection benchmarks [6]. We refer to it as **classification+localization accuracy** (or **cls+loc** for short). We also at times evaluate just **classification accuracy** (or **cls** for short), which is the fraction of images on which the correct class label C was matched by one of the five proposed labels, without considering whether the instance was correctly localized.

2.2. Quantitative measures of localization difficulty

We introduce three metrics of localization difficulty: number of object instances per image, chance performance of localization (closely related to average object scale), and the level of clutter. These serves multiple purposes throughout the paper: (1) to compare the ILSVRC dataset to the standard PASCAL benchmark in this section,[3] (2) to evaluate the accuracy of different detection algorithms as a function of these measures later in Section 3.2, and (3) to control for these image statistics when evaluating the accuracy of detection algorithms as a function of inherent object properties such as real-world object size later in Section 3.3.

Measures of localizaiton difficulty are computed on the validation set of both datasets.

Instances per image. Real-world scenes are likely to contain multiple instances of some objects, and nearby object instances are particularly difficult to delineate and localize. The average object category in ILSVRC has 1.62 target object instances on average per image, with each instance having 0.47 neighbors (adjacent instances of the same object

[2] Please see supplement for analysis of top-5 evaluation criteria.

[3] Please see supplement for more ILSVRC/PASCAL comparisons.

category) on average. This is comparable to 1.69 instances per image and 0.52 neighbors in PASCAL.

Chance performance of localization (CPL). Chance performance on a dataset is a common metric to consider. We define the CPL measure as the expected accuracy of a detector which first randomly samples an object instance of that class and then uses its bounding box directly as the proposed localization window on all other images (after rescaling the images to the same size). Concretely, let $B_1, B_2, \ldots, B_N$ be all the bounding boxes of the object instances within a class, then

$$\text{CPL} = \frac{\sum_i \sum_{j \neq i} IOU(B_i, B_j) \geq 0.5}{N(N-1)} \qquad (1)$$

Some of the most difficult ILSVRC categories to localize according to this metric are basketball, swimming trunks, ping pong ball and rubber eraser, all with less than 0.2% CPL. This measure correlates strongly ($\rho = 0.9$) with the average scale of the object (fraction of image occupied by object). The average CPL across the 1000 ILSVRC categories is 20.8%. The 20 PASCAL categories have an average CPL of 8.7%, which is the same as the CPL of the 562 most difficult categories of ILSVRC.

Clutter. Intuitively, even small objects are easy to localize on a plain background. To quantify clutter we employ the objectness measure of [1], which is a class-generic object detector evaluating how likely a window in the image contains a coherent object (of any class) as opposed to background (sky, water, grass). For every image m containing target object instances at positions $B_1^m, B_2^m, \ldots$, we use the publicly available objectness software to sample 1000 windows $W_1^m, W_2^m, \ldots W_{1000}^m$, in order of decreasing probability of the window containing any generic object. Let OBJ(m) be the number of generic object-looking windows sampled before localizing an instance of the target category, i.e., $\text{OBJ}(m) = \min\{k : \max_i \text{IOU}(W_k^m, B_i^m) \geq 0.5\}$. For a category containing M images, we compute the average number of such windows per image and define

$$\text{CLUTTER} = \log_2(\tfrac{1}{M} \textstyle\sum_m \text{OBJ}(m)) \qquad (2)$$

The higher the clutter of a category, the harder the objects are to localize according to generic cues. If an object can't be localized with the first 1000 windows (as is the case for 1% of images on average per category in ILSVRC and 5% in PASCAL), we set $\text{OBJ}(m) = 1001$. The fact that more than 95% of objects can be localized with these windows imply that the objectness cue is already quite strong, so objects that require many windows on average will be extremely difficult to detect: e.g., ping pong ball (clutter of 9.57, or 758 windows on average), basketball (clutter of 9.21), puck (clutter of 9.17) in ILSVRC. The most difficult object in PASCAL is bottle with clutter score of 8.47. On average, ILSVRC has clutter score of 3.59. The most difficult subset of ILSVRC with 250 object categories has an order of magnitude more categories and the same average amount of clutter (of 5.90) as the PASCAL dataset.

2.3. State-of-the-art detection algorithms

Several object detection systems participated in the ImageNet classification+localization challenge of 2012; the winners are good candidates for our analysis. Using two very distinct leading algorithms, we analyze the current successes and weaknesses in large-scale detection.

The analysis in this paper requires only per-class accuracies and class confusion matrices of these methods.

SV system. The winning system of the challenge, named SuperVision and abbreviated as SV, uses neural networks to learn the full image representation automatically from data. It is based upon a supervised convolutional neural network with 7 hidden layers, trained using stochastic gradient descent on the GPU [12]. This system was targeted to image classification so its strong performance on object localization is particularly impressive.

VGG system. The other algorithm, OXFORD_VGG, is based on the more conventional image classification and detection pipeline. It uses an image classification system with dense SIFT features and color statistics [13], a Fisher vector representation [16], and a linear SVM classifier, plus additional insights from [2, 17], combined with the deformable parts-based model (DPM) [9] which has been the dominant model for generic object detection for many years.

Upper bound. We also consider an optimistic upper bound which combines the outputs of the VGG and SV on every image. Here the output on the image is considered correct if any of the 10 predicted (class, location) pairs are correct. Evaluating this joint algorithm helps to summarize the common trends of SV and VGG as well as to illustrate the key scenarios where SV and VGG provide complementary sources of information.[4]

3. Where are we as a field?

We present our analysis of the state of categorical object detection on an unprecedented scale of 1000 object categories using the large-scale ILSVRC dataset.

3.1. How accurate are the current algorithms?

Object detection involves both correctly predicting the class label and localizing the object; in seeking to understand the limitations of current algorithms we initially decouple these two measures. The hierarchical structure of ILSVRC allows us to analyze the accuracy of the algorithm as a function of semantic depth; in other words, we can relax the requirement that a Dalmatian be classified as a Dal-

[4] Please see supplement for discussion and analysis of this upper bound.

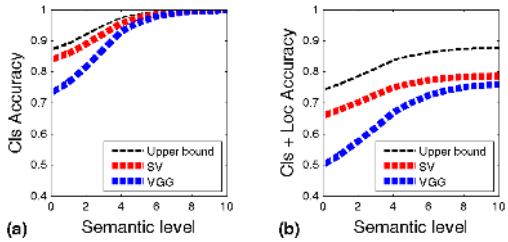

Figure 3. (a) Classification and (b) classification with localization accuracy as a function of moving up the ImageNet hierarchy. [4] At level 0 the algorithm is required to produce the exact class label; at level 1 any sister label is accepted, at level 10 any leaf node of generic concepts such as "living thing" is accepted.

matian and instead evaluate the accuracy of the algorithm when allowing any dog breed to be an acceptable label. In Figure 3 we plot the cls and cls+loc accuracies of methods described in Section 2.3 as a function of semantic depth: at level 0 we require the exact label, at level 1 we accept any sister concept (called *synset* in ImageNet [4]), and so forth. As we move up the hierarchy errors made from fine-grained classification are corrected. For reference, breeds of dogs are on average 4 levels removed from the "domestic dog" node and birds are 4.6 levels removed from "bird" node.

Three things are apparent from this analysis: (1) basic-level classification is surprisingly accurate: SV distinguishes dogs from other objects with > 0.99 accuracy, and birds with 0.98 accuracy; (2) despite this, the gap between cls and cls+loc accuracies remains large (> 0.180 at all semantic levels); (3) SV and VGG provide complementary sources of information on cls+loc: the upper bound is 0.083 higher than either SV or VGG alone.[5] The last two observations prompt more detailed investigation.

3.2. How do image-level statistics affect localization accuracy?

Despite high levels of classification accuracy on basic-level categories, localization is far from perfect. To better understand what scenarios are challenging for current detection algorithms, we begin by evaluating the accuracy as a function of global image statistics. Figure 4 visualizes the cls+loc accuracy of the detection algorithms described in Section 2.3 as a function of the quantitative measures of localization difficulty introduced in Section 2.2. We analyze these measures one by one.

Instances per image. SV is more strongly correlated with the number of instances of the object than VGG (correlation of -0.436 versus -0.052). *This suggests that SV would be good at, e.g., noticing that there are cars in the image but might have difficulty separating out two nearby cars.* On categories with more than 2 instances per image, the correlation between number of instances and SV accuracy is significantly weaker (-0.156), implying that multiple objects are challenging regardless of the exact number. Combining the global model of SV with the strong object boundary model of VGG is a promising future direction.

Chance Performance of Localization (CPL). The correlation of CPL with SV accuracy is double that with VGG (0.640 vs 0.315) and the slope of the regression line is two times steeper (0.781 of SV, 0.356 of VGG) so SV's accuracy degrades faster and more consistently as CPL get smaller. In fact, when considering 225 object categories with lowest CPL the cls+loc accuracy of SV and VGG is the same at 0.404, and on smaller objects VGG outperforms SV.[6] CPL is highly correlated with object scale as noted in Section 2.2. *VGG is better suited than SV for localizing small objects.*

Some categories contain a bimodal distribution of images accounting for low CPL: a mixture of close-ups with one object instance occupying the whole image and images depicting a large cluster of small objects. *SV and VGG algorithms perform well on different subsets of such data and combine to create a much stronger detector.* Some example categories are screw (CPL 0.9%, upper bound accuracy 0.624, SV accuracy 0.269, VGG accuracy 0.398) or boathouse (CPL 1.1%, upper bound accuracy 0.691, SV accuracy 0.443, VGG accuracy 0.454).

Clutter. The measure of clutter is defined in Section 2.2 as a logarithm of the number of objectness windows [1] it takes on average to localize an instance of the target class. There is a strong correlation between this measure and the accuracy of all the algorithms (VGG correlation -0.445, SV -0.704). *This shows that clutter may be a useful metric for evaluating the difficulty of detection datasets.*[7]

With this in mind, we consider a subset of 250 ILSVRC categories which has the same average clutter of 5.90 as the PASCAL dataset. On this subset, SV achieves cls+loc accuracy of 0.439, still significantly outperforming VGG with accuracy of 0.374.[8]

3.3. What intrinsic properties of object categories affect localization accuracy?

So far we have studied how image-level properties affect detection, and now turn to examine the effects of intrinsic object properties; access to results for up to a thousand categories allows us to perform this analysis. We first

[5] An interesting side note is that the probability of correctly localizing the object given that it was correctly classified remains the same for SV at all levels of the semantic hierarchy (at 78%) but increases dramatically for VGG (from 68% to 76%). This confirms the intuition that SV is well suited for fine-grained classification while VGG is best used for localizing basic object categories.

[6] Please see supplement for plot of CPL versus accuracy of algorithms.

[7] Additionally, this implies that the number of objectness windows [1] is logarithmically related to the difficulty of localization.

[8] A similar conclusion holds true with the CPL metric as well. On the 562 hardest ILSVRC categories with the same average CPL as the PASCAL categories, SV achieves cls+loc accuracy of 0.554, significantly outperforming VGG with cls+loc accuracy of 0.461.

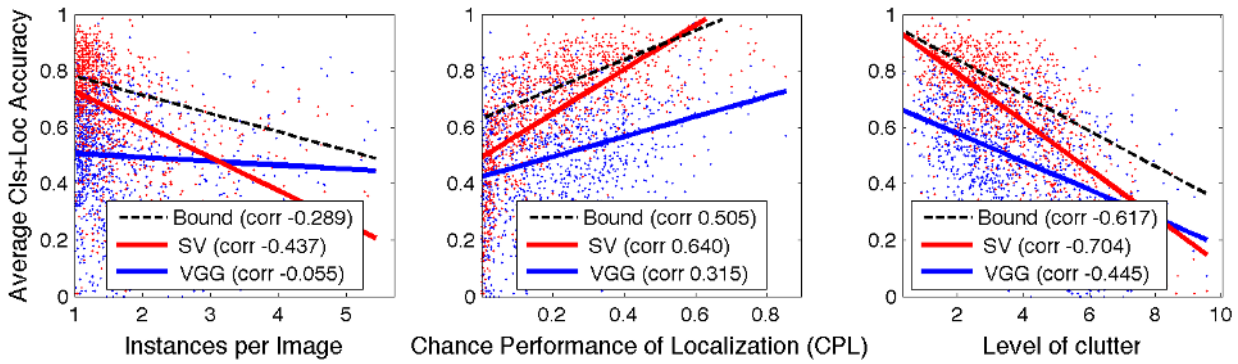

Figure 4. The impact of several quantitative measures of localization difficulty (Section 2.2) on the cls+loc accuracy. Each dot corresponds to one of the 1000 object categories of ILSVRC2012 and to one of the two algorithms (SV in red, VGG in blue). The difficulty measure (x-axis) is computed on the validation set; the accuracy of the algorithm (y-axis) is evaluated on the test set. The best fit linear models for each algorithm are also shown to summarize trends. The black line corresponds to the upper bound combination of SV and VGG; the accuracy on individual class of this method is not shown to reduce clutter. Please refer to Section 3.2 for analysis.

consider the differences between natural and man-made objects. Next we define five additional properties inspired by human vision: real-world size, deformability within instance, amount of texture, distinctiveness of color and distinctiveness of shape, all visualized in Figure 1. Human subjects annotated each of the 1000 object categories with these properties; we specify the domains of these properties below as we discuss each one. Figure 5 visualizes the effects of these properties on the performance of object detectors, and here we summarize the key findings.

Natural vs man-made objects: natural (e.g., bee, scuba diver, lemon) or man-made (e.g., sock, trolley, apron)

The accuracy of object detectors is strongly correlated with whether the object is natural or man-made. Figure 5(a) shows that SV achieves cls+loc accuracy of 0.768 on the 427 natural ILSVRC classes compared to only 0.570 on the 573 man-made classes; VGG achieves 0.514 on natural compared on 0.486 on man-made.

Importantly, this observation holds regardless of the image-level statistics of Section 3.2. Figure 5(b) shows the cumulative loc+cls accuracy as a function of increasing CPL on natural and man-made objects separately. Regardless of how difficult of a subset of ILSVRC is chosen, the accuracy of SV on natural objects is always at least 0.186 higher than its accuracy of man-made objects. Interestingly, on the more challenging categories (e.g., the 330 categories with CPL $\leq$ 0.1), VGG's cls+loc accuracy on natural objects is also 0.143 higher than on man-made ones even though on average over 1000 categories VGG's accuracy is not as strongly affected by this property.[9]

Real-world size: tiny (e.g. nail), small (e.g. fox), medium (e.g. bookcase), large (e.g. car), huge (e.g. church).

Algorithms tend to perform better on objects which are larger in the real-world, both on natural objects (Figure 5(c)) and on man-made objects (Figure 5(d)). One exception is huge man-made objects which are slightly more difficult to localize than the man-made large objects: since many huge objects are buildings (e.g., church, shoe shop, prison) they may not be fully visible in the image or the picture might be taken inside the building, so localizing them may be particularly challenging.[10]

Tiny objects are very challenging for both SV and VGG; however, combining the algorithms yields significant improvement. The cls+loc accuracy of the upper bound method is 0.111 higher than the accuracy of either SV or VGG on tiny objects. This is significantly better than improvements of 0.071, 0.077, 0.090 and 0.072 on the small, medium, large and huge objects respectively.

Distinctiveness of color: none (e.g. clothes), low (e.g. cleaver), medium (e.g. hay), high (e.g. tennis ball).

Distinctiveness of color makes it easier to detect objects; however, the benefits of color distinctiveness are not as pronounced when controlling for whether the object is natural or man-made. Figure 5(e) shows that objects which have distinctive colors tend to be easier to classify and localize than those that don't. The upper bound cls+loc accuracy is 0.117 higher on objects with distinctive color than on those without.[11] Since color distinctiveness is strongly correlated with being man-made (84% of natural objects have distinctive color compared to only 21% of man-made objects), we evaluate on man-made versus natural objects separately. Restricted to man-made objects, cls+loc accuracy of upper bound is only 0.044 higher on objects with distinctive color than on those without; restricted to natural objects, it is only 0.040 higher for objects distinctive in color.

[9] Similar patterns emerge with clutter measure instead of CPL.

[10] There are only a few huge *natural* objects so we don't analyze them.

[11] There was significant variability in the human annotations for this attribute (color distinctiveness is difficult to precisely define), so we simplify to two cases: has distinctive color ("medium" or "high") or does not.

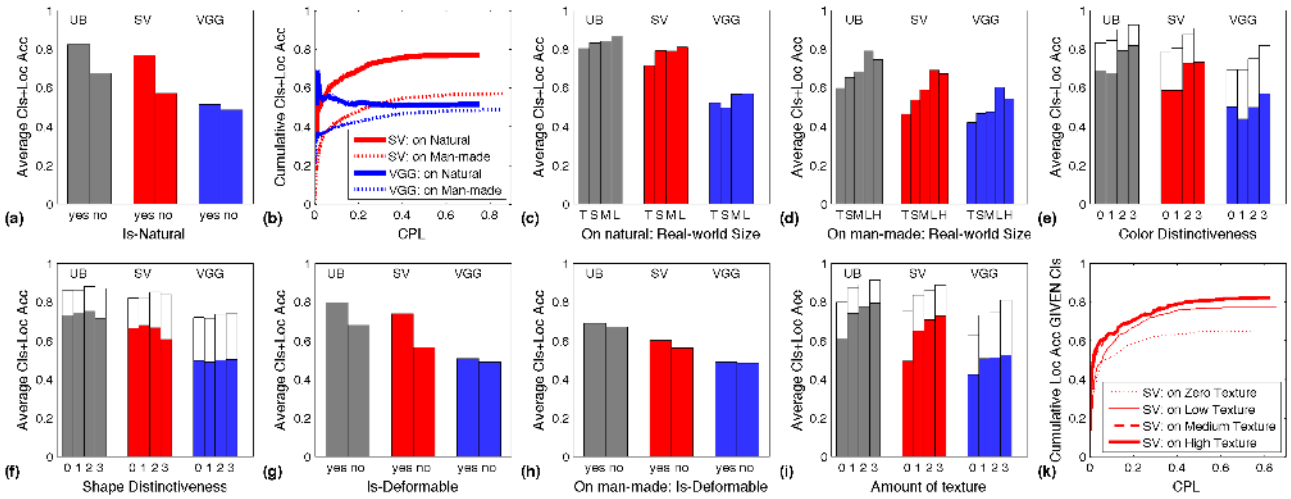

Figure 5. Cls+loc accuracy of Upper Bound (gray), SV (red) and VGG (blue) as a function of the intrinsic properties of the ILSVRC2012 object categories. Plots except (b,k) show the average cls+loc accuracy of the algorithms on subsets of ILSVRC. Plots (e,f,i) additionally report classification accuracy (white bars). Plot (b) shows the cumulative cls+loc accuracy as a function of CPL (Eq. 1). The height of the curve corresponds to the average accuracy of the object categories with equal or smaller CPL measures. Plot (k) is similar to (b) except the y-axis corresponds to the cls+loc accuracy *only on images which were correctly classified*. Please see Section 3.3 for analysis.

Distinctiveness of shape: none (e.g. chocolate sauce), low (e.g. tape player), medium (e.g. T-shirt), high (e.g. banana).

Figure 5(f) visualizes average cls and cls+loc accuracies of the three algorithms across the different levels of shape distinctiveness. *There is no observed correlation between human-annotated distinctiveness of shape and the accuracy of the algorithms* (this holds true when considering subsets of the object classes, such as just man-made objects as well). This is consistent with the intuition that general object category detection algorithms tend to avoid rigid modeling of object shape and instead rely on other cues instead, such as texture or color.

Deformability within instance: yes (e.g., water snake) or no (e.g., mug).

One aspect of object shape that is often modeled is deformability within instance. [9] *Whether or not the object is deformable has relatively little bearing on the performance of the algorithms in the absence of other factors.* VGG is largely unaffected by deformability: cls+loc accuracy of 0.507 on deformable versus 0.489 on non-deformable objects as shown in Figure 5(g). On average SV is significantly more accurate on deformable objects than non-deformable ones (cls+loc accuracy 0.740 versus 0.566); however, when evaluating separately just on the 573 man-made classes in Figure 5(h) the effect becomes significantly less pronounced (cls+loc accuracy 0.602 on 116 deformable objects versus 0.566 on 457 non-deformable ones).[12]

Amount of texture: zero (e.g. punching bag), low (e.g. horse), medium (e.g. sheep), high (e.g. honeycomb).

Both algorithms are much more accurate on textured objects in both classification and classification with localization as shown in Figure 5(i)).[13]

We now consider just the images within each class which were correctly classified by VGG and compute the per-class localization accuracy conditioned on correct classification. Grouping the classes across the four levels of texture, we observe that VGG correctly localizes the object in between $65-68\%$ of the correctly classified images on average per class for every level. For SV, however, the pattern is different: on untextured objects SV also accurately localizes 65% of the correctly classified images on average, but on highly textured it localized 82%! In Figure 5(k) we show that this pattern holds even across different CPL.[14] *Once the image has been correctly classified, SV correctly localizes many more of the textured than untextured objects.*

4. Discussion: what have we learned and where should we be going?

We summarize several key observations from the above analysis which provide guidance for future research.

ILSVRC as a benchmark dataset. We demonstrate that the ILSVR dataset provides many of the same localization challenges as the PASCAL dataset, especially when con-

[12] The set of 427 natural classes contains only 32 non-deformable objects (e.g., strawberry, lemon, rose hip), so we omit it from the analysis.

[13] A similar pattern appears on both man-made and natural classes independently. Man-made objects have average texture of 2.2 while natural objects are slightly more textured at 2.8.

[14] Level of texture is correlated with CPL (going from zero to high texture, the average CPL is 12.7, 19.9, 23.7, and 26.4).

sidering smaller subsets such as the 562 most difficult categories according to CPL or 250 more difficult categories according to level of clutter. As a result of its scale, the ILSVRC allows for evaluation of many other object detector properties which may be important for real-world applications, such as the ability of detectors to differentiate fine-grained object categories, and also allows for evaluating the performance of detectors under a variety of image-level statistics and object class properties. This is in line with the goal of [10] to perform deeper analysis of detectors instead of reporting a single accuracy metric.

Focus on fine-grained recognition. Classification accuracy of basic-level categories is already quite high; however, distinguishing between more fine-grained classes is much more challenging for current methods. Consistent with the recent trend in recognition literature, this reinforces the need to focus on capturing fine-grained distinction between classes in seeking to better understand the visual world.

Clutter measure for evaluating datasets and targeting algorithms. The measure of clutter using the latest techniques in unsupervised object discovery [1] defined in Section 2.2 is strongly correlated with the accuracy of current state-of-the-art algorithms. While the high-level insight that clutter is detrimental to the performance of detection algorithms is not novel, this suggests that the introduced metric is useful to consider when collecting new datasets or designing the next generation of object detectors.

Combining detection algorithms. The current leading detectors SV and VGG are found to be complementary to each other on categories with low CPL in Section 3.2 and, similarly, on objects which are "tiny" in the real world in Section 3.3. This may be intuitive given prior knowledge about the design of the SV and VGG systems, but it is still useful to quantify. This is a key domain to consider when designing the next generation of detectors combining the benefits of the current leading systems.

Acknowledgements

Much of the credit goes to the developers of the state-of-the-art detection systems: A. Krizhevsky, I. Sutskever, G. Hinton (SuperVision) and K. Simonyan, Y. Aytar, A. Vedaldi and A. Zisserman (Oxford_VGG). This work was in part supported by Google.

References

[1] B. Alexe, T. Deselares, and V. Ferrari. Measuring the objectness of image windows. In *PAMI*, 2012. 4, 5, 8

[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 4

[3] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. Large Scale Visual Recognition Challenge. http://www.image-net.org/challenges/LSVRC/2012/, 2012. 1, 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *CVPR*, 2009. 1, 3, 5

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

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 2, 3

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, June 2010. 1, 2, 3

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few examples: an incremental bayesian approach tested on 101 object categories. In *CVPR*, 2004. 1

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32, 2010. 1, 2, 4, 7

[10] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 1, 2, 3, 8

[11] D. Hoiem, C. Rother, and J. Winn. 3D layout CRF for multi-view object class recognition and segmentation. In *CVPR*, 2007. 2

[12] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2, 4

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 4

[14] N. Pinto, D. Cox, and J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Comp Biology*, 4, 2008. 2

[15] B. Russell, A. Torralba, K. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 2007. 1

[16] J. Sanchez and F. Perronnin. High-dim. signature compression for large-scale image classification. In *CVPR*, 2011. 4

[17] J. Sanchez, F. Perronnin, and T. de Campos. Modeling spatial layout of images beyond spatial pyramids. In *PRL*, 2012. 4

[18] A. Torralba and A. Efros. An unbiased look at dataset bias. In *CVPR*, 2011. 1, 2

[19] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. In *PAMI*, 2008. 1

[20] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 1

[21] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009. 1

[22] X. Wang, T. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009. 2

[23] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005. 2

[24] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from Abbey to Zoo. *CVPR*, 2010. 1

[25] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes. Do we need more training data or better models for object detection. In *BMVC*, 2012. 2