Salient Object Subitizing

Jianming Zhang[1] Shugao Ma[1] Mehrnoosh Sameki[1] Stan Sclaroff[1] Margrit Betke[1]
Zhe Lin[2] Xiaohui Shen[2] Brian Price[2] Radomír Měch[2]
[1]Boston University [2]Adobe Research

Abstract

People can immediately and precisely identify that an image contains 1, 2, 3 or 4 items by a simple glance. The phenomenon, known as Subitizing, inspires us to pursue the task of Salient Object Subitizing (SOS), i.e. *predicting the existence and the number of salient objects in a scene using holistic cues. To study this problem, we propose a new image dataset annotated using an online crowdsourcing marketplace. We show that a proposed subitizing technique using an end-to-end Convolutional Neural Network (CNN) model achieves significantly better than chance performance in matching human labels on our dataset. It attains 94% accuracy in detecting the existence of salient objects, and 42-82% accuracy (chance is 20%) in predicting the number of salient objects (1, 2, 3, and 4+), without resorting to any object localization process. Finally, we demonstrate the usefulness of the proposed subitizing technique in two computer vision applications: salient object detection and object proposal.*

1. Introduction

How quickly can you tell the number of **salient** objects in each image in Fig. 1? It was found over a century ago that people are equipped with a remarkable capacity to effortlessly and consistently identify 1, 2, 3 or 4 items by a simple glance [27]. This phenomenon, later coined by Kaufman, et al. as *Subitizing* [29], has been observed under various measurements [6, 37]. It is shown that apprehension of small numbers up to three or four is highly accurate, quick and confident, while beyond this subitizing range, the feeling is lost. Accumulating evidence also shows that infants and even certain species of animals can differentiate between small numbers of items within the subitizing range [18, 25, 17, 40], suggesting that subitizing may be an inborn numeric capacity of humans and animals. It is speculated that subitizing is a preattentive and parallel process [18, 52, 54], and that it can help humans and animals make prompt decisions in basic tasks like navigation, searching and choice making [42, 24].

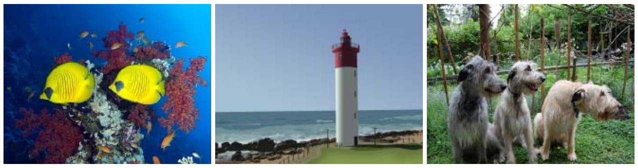


Figure 1: How fast can you tell the number of prominent objects in each of these images?

In this paper, we propose a subitizing-like approach to estimate the number (0, 1, 2, 3 and 4+) of salient objects in a scene, without resorting to any object localization process. Solving this *Salient Object Subitizing* (SOS) problem can benefit many computer vision tasks and applications.

Knowing the existence and the number of salient objects without the expensive detection process (*e.g.*, sliding window detection) can enable a machine vision system to select different processing pipelines at an early stage, making it more intelligent and reducing computational cost. For example, SOS can help a computer vision system suppress the object detection process, until the existence of salient objects is detected, and it can also provide cues for selecting among search strategies and early stopping criteria based on the predicted number. Differentiating between scenes with zero, a single and multiple salient objects can also facilitate applications like robot vision [45], egocentric video summarization [31], snap point prediction [57], iconic image detection [7] and image thumbnailing [14], *etc.*

To study the SOS problem, we provide a new image dataset of about 7000 images, where the number of salient objects in each image has been annotated by Amazon Mechanical Turk (AMT) workers. In Fig. 2, we show some sample images in the proposed SOS dataset with the collected ground-truth labels. Although there are no bounding box annotations accompanying the numbers, it is usually pretty straightforward to see which objects these numbers refer to. The annotations from the AMT workers are further analyzed in a more controlled offline setting, which shows a high inter-subject consistency in subitizing salient objects.

Our ultimate goal is to develop a fast and accurate computational method to estimate the number of salient objects

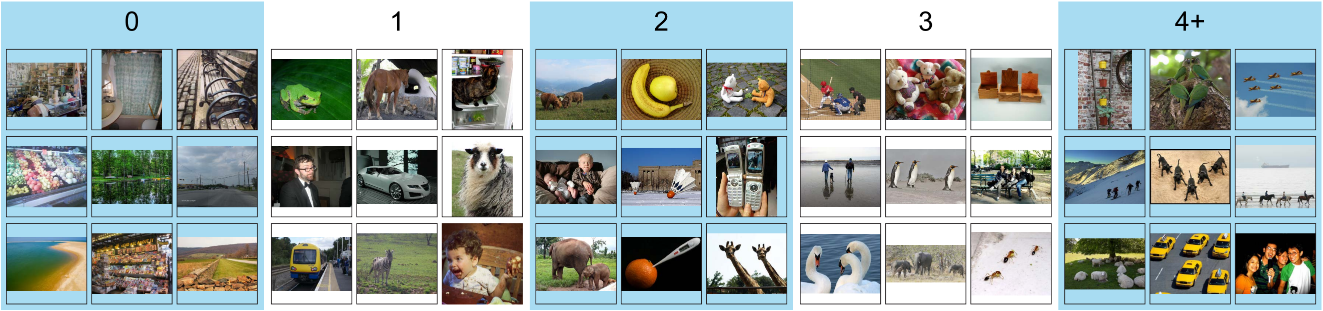


Figure 2: Sample images of the proposed SOS dataset. These images cover a wide range of content and object categories.

in natural images. The trivial counting-by-detection approach is quite challenging in this scenario, due to cluttered background, occlusion, and large appearance, position and scale variations of the objects in everyday images (see our collected images in Fig. 2). Instead, inspired by the psychological observation that the subitizing is likely to be accomplished by recognizing holistic patterns [26, 37, 15, 9], the proposed SOS method bypasses the challenging salient object localization process by using global features.

An implementation of our SOS method using an end-to-end Convolutional Neural Network (CNN) classifier attains 94% accuracy in detecting the existence of salient objects, and 42-82% accuracy (chance is 20%) in predicting the number of salient objects (1, 2, 3, and 4+) on our dataset, without resorting to any intermediate saliency map computation or salient object detection. These results are quite encouraging, considering that our CNN-based SOS method is capable of processing an image in a couple of milliseconds.

To summarize, the key contributions of this paper are:

1. We formulate the Salient Object Subitizing (SOS) problem, which aims to predict the number of salient objects in a scene without resorting to any object localization process.

2. We provide an annotated benchmark dataset for evaluation of SOS methods.

3. We present a simple CNN-based implementation of SOS, which achieves promising results, while being capable of processing an image in a few milliseconds.

4. We demonstrate applications of the SOS technique in guiding salient object detection and object proposal generation, resulting in state-of-the-art performance.

In the task of salient object detection [35, 48], we demonstrate that SOS can help improve accuracy by identifying images that contain no salient object. In the task of object proposal generation [59, 2], we present a simple content-aware proposal allocation approach using SOS, and show consistent improvement over state-of-the-art.

2. Related Work

Salient object detection. Salient object detection aims at localizing salient objects in a scene by a foreground mask [1, 13] or bounding boxes [35, 23, 21, 48]. However, existing salient object detection methods assume the existence of salient objects in an image. Furthermore, those methods are often optimized for images that contain a single salient object [33, 8]. Thus, counting salient objects using existent salient object detection methods can be quite unreliable.

Detecting the existence of salient objects. Some works address the problem of detecting the existence of salient objects in an image. In [55], a global feature based on several saliency maps is used to determine the existence of salient objects in thumbnail images, assuming an image either contains a single salient object or none. In [45], saliency histogram features are exploited to detect the existence of interesting objects for robot vision. It is worth noting that the testing images in [55, 45] are substantially simplified compared to ours, and the methods of [55, 45] cannot provide information about the number of salient objects.

Automated object counting. There is a large body of literature about object counting based on density estimation [32, 5], object detection/segmentation [50, 38, 3] and regression [10, 11]. These works usually rely on category-dependent training, and assume that the target objects have similar appearances and sizes. The proposed SOS problem is quite different from these counting approaches, in that it targets at category-independent inference of the number of generic salient objects, whose appearance can dramatically vary from category to category, and from image to image.

Modeling visual numerosity. Some researchers exploit deep neural network models to analyze the emergence of visual numerosity in human and animals [49, 60]. In these works, abstract binary patterns are used as training data, and the researchers study how the deep neural network model captures the number sense during either unsupervised or supervised learning. Our work looks at a more application-oriented problem, and targets at inferring the number of salient objects in natural images.

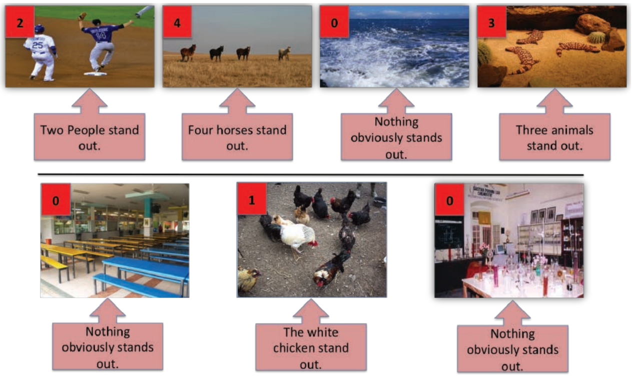

2
Two People stand out.
4
Four horses stand out.
0
Nothing obviously stands out.
3
Three animals stand out.
0
Nothing obviously stands out.
1
The white chicken stand out.
0
Nothing obviously stands out.

Figure 3: Example labeled images for AMT workers. The number of salient objects is shown in the red rectangle on each image. There is a brief explanation below each image.

3. The SOS Dataset

We describe the collection of the Salient Object Subitizing dataset, and then provide the labeling consistency analysis of the annotation collected via Amazon Mechanical Turk. The dataset is available on our project website[1].

3.1. Image Source

To collect a dataset of images with different numbers of salient objects, we gathered a set of images from three object detection datasets, COCO [34], ImageNet [44] and VOC07 [19], and a scene dataset, SUN [56]. This preliminary set is composed of about 17000 images in total. 2000 images are from the SUN dataset, and about 5000 images are from each of the other three datasets.

For VOC07, the whole train and validation set is included. We limit the number of images from the SUN dataset to 2000, because most images in this dataset do not contain obviously salient objects, and we do not want the images from this dataset to dominate the category for no salient object. The 2000 images are randomly sampled from SUN. For the COCO and ImageNet dataset[2], we use the bounding box annotations to split the dataset into four categories for 1, 2, 3 and 4+, and then sample an equal number of images from each category, in the hope that this can help balance the distribution of our final dataset.

3.2. Annotation Collection

We used the crowdsourcing platform Amazon Mechanical Turk (AMT) to collect annotations for our preliminary set of images. We asked the AMT workers to label each image as containing 0, 1, 2, 3 or 4+ prominent objects. Several example labeled images (shown in Fig. 3) were provided prior to each task as an instruction. We purposely did not give more specific instructions regarding some amibiguous cases for counting, *e.g*. counting a man riding a horse as one or two objects. Each task, or HIT (Human Intelligence Task) was composed of five images with a 2-minute time limit, and the compensation was one cent per task. All five images in one task were displayed at the same time. The average completion time for each task was about 20s. We collected five annotations per image from distinct workers. About 260 distinct workers contributed to this dataset, and 90% of the tasks were completed by 60 workers.

Table 1: Distribution of images in the SOS dataset

category	COCO	VOC07	ImageNet	SUN	total
0	271	184	233	943	1631
1	1236	1388	478	88	3190
2	314	376	272	34	996
3	92	74	555	17	738
4+	151	119	72	3	345
total	2064	2141	1610	1085	6900

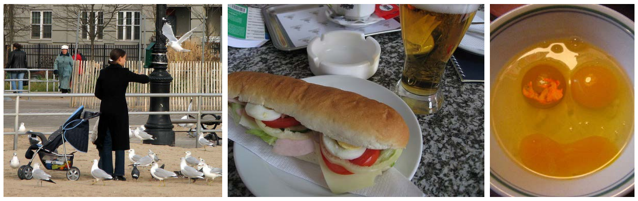


Figure 4: Sample images with divergent labels. These images are a bit ambiguous about what should be counted as an individual salient object.

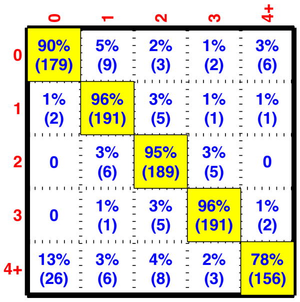
A few images do not have a clear notion about what should be counted as an individual salient object, and labels on those images tend to be divergent. We show some of these images in Fig. 4. We exclude images with fewer than four consensus labels, leaving 6900 images for our final SOS dataset. In Table 1, we show the joint distribution of images with respect to the labeled category and the original dataset. The category distributions of the images from COCO and VOC07 are very similar, and the majority of the images from the SUN dataset belong to the "0" category. The ImageNet dataset contains more images with three salient objects than the other datasets.

3.3. Annotation Consistency Analysis

During the annotation collection process, we simplified the task for the AMT workers by giving them 2 mins to label five images at a time. This simplification allowed us to gather a large number of annotations with reduced time and cost. However, the flexible viewing time allowed the AMT workers to look closely at these images, which may have had an influence over their attention and their answers to the number of salient objects. This leaves us with a couple im-

[1] http://www.cs.bu.edu/groups/ivc/Subitizing/

[2] We use the subset of ImageNet images with bounding box annotations.



	0	1	2	3	4+
0	90% (179)	5% (9)	2% (3)	1% (2)	3% (6)
1	1% (2)	96% (191)	3% (5)	1% (1)	1% (1)
2	0	3% (6)	95% (189)	3% (5)	0
3	0	1% (1)	3% (5)	96% (191)	1% (2)
4+	13% (26)	3% (6)	4% (8)	2% (3)	78% (156)

Figure 5: Averaged confusion matrix of our offline human subitizing test. Each row corresponds to a ground-truth category labeled by AMT workers. The percentage reported in each cell is the average proportion of images of the category A (row number) labeled as category B (column number). For over 90% images, the labels from the offline subitizing test are consistent with the labels from AMT workers.

Table 2: Human subitizing accuracy in matching category labels from Mechanical Turk workers.

	sbj.1	sbj.2	sbj.3	Avg.
Accuracy	90%	92%	90%	91%

portant questions. Given a shorter viewing time, will labeling consistency among different subjects decrease? Moreover, will shortening the viewing time change the common answers to the number of salient objects? Answering these question is critical in understanding our problem and data.

To answer these questions, we conducted a more controlled offline experiment based on common experimental settings in subitizing literature [6, 37]. In this experiment, only one image was shown to a subject at a time, and this image was exposed to the subject for only 500 ms. After that, the subject was asked to tell the number of salient objects by choosing an answer from 0, 1, 2, 3, and 4+.

We randomly selected 200 images from each category according to the labels collected from AMT. Three subjects were recruited for this experiment, and each of them was asked to complete the labeling of all 1000 images. We divided that task into 40 sessions, each of which was composed of 25 images. The subjects received the same instructions as the AMT workers, except they were exposed to one image at a time for 500 ms. Again, we intentionally omitted specific instructions for ambiguous cases for counting.

Over 98% test images receive at least two out of three consensus labels in our experiment, and all three subjects agree on 84% of the test images. Table 2 shows the proportion of category labels from each subject that match the labels from AMT workers. All subjects agree with AMT workers on over 90% of sampled images. To see details of the labeling consistency, we show in Fig. 5 the averaged confusion matrix of the three subjects. Each row corresponds to a category label from the AMT workers, and in each cell, we show the average number (in the brackets) and percentage of images of category A (row number) classified as category B (column number). For categories 1, 2 and 3, the per-class accuracy scores are above 95%, showing that limiting the viewing time has little effect on the answers in these categories. For category 0, there is a 90% agreement between the labels from AMT workers and from the offline subitizing test, indicating that changing the viewing time may slightly affect the apprehension of salient objects. For category 4+, there is only 78% agreement, and about 13% of images in this category are classified as category 0.

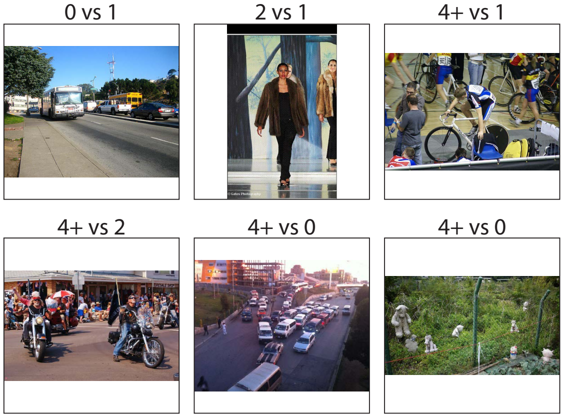


Figure 6: Sample images that are consistently labeled by all three subjects in our offline subitizing test as a different category from what is labeled by the Mechanical Turk workers. Above each image, there is the AMT workers' label (left) vs the offline-subitizing label (right).

In Fig. 6, we show sample images that are consistently labeled by all three subjects in our offline subitizing test as a different category than labeled by AMT workers. We find some labeling discrepancy may be attributed to the fact that objects at the image center tend to be thought of as more salient than other ones given a short viewing time (see images in the top row of Fig. 6). In addition, some images with many foreground objects (far above the subitizing limit of 4) are labeled as 4+ by AMT workers, but they tend to be labeled as category 0 in our offline subitizing test (see the middle and right images at the bottom row in Fig. 6).

Despite the labeling discrepancy on a small proportion of the sampled images, limiting the viewing time to a fraction of a second does not significantly decrease the inter-subject consistency or change the answers to the number of salient objects on most test images. We thereby believe the proposed SOS dataset is valid. The per-class accuracy shown in Fig. 5 (percentage numbers in diagonal cells) can be interpreted as an estimate of the human performance baseline on our dataset.

4. Salient Object Subitizing

Since it remains an open problem to robustly detect salient objects, we propose a Salient Object Subitizing method for estimating the number of salient objects without resorting to any object detection process. We benchmark several simple implementations of the SOS method to demonstrate the value of this problem.

4.1. Global Image Features

Although salient objects can have dramatically different appearance in color, texture and shape, we expect that global geometric information can be used to differentiate images with different numbers of salient objects. Therefore, we evaluate HOG [16], GIST [51] and Improved Fisher Vectors (IFV) [41] of dense SIFT [36], all of which are gradient-based features and have been used for image classification. We also evaluate a spatial pyramid feature of saliency maps, in the hope that saliency maps may provide information about the composition of the foreground. Finally, inspired by the remarkable progress made by Convolutional Neural Network (CNN) features in many computer vision problems [22, 30, 43, 46], we try the CNN feature. In [22], it is suggested that given limited domain specific data, fine-tuning a pre-trained CNN model can be an effective and highly practical approach for many problems. Thus, we fine-tune the pre-trained CNN model of [30] for our problem.

The implementation details for each of the global feature representations are as follows:

GIST. The GIST descriptor [51] is computed based on 32 Gabor-like filters with varying scales and orientations. We use the implementation of [51] to extract a 512-D GIST feature, which is a concatenation of averaged filter responses over a 4×4 grid.

HOG. We use the implementation by [20] to compute HOG features. Images are first resized to 128×128, and HOG descriptors are computed on a 16×16 grid, with the cell size being 8×8. The HOG features of image cells are concatenated into a 7936-D feature. We have also tried combining HOG features computed on multi-scale versions of the input image, but this gives little improvement.

IFV. We use the implementation by [12]. The codebook size is 256, and the dimensionality of SIFT descriptors is reduced to 80 by PCA. Hellinger's kernel and L2-normalization is used with this encoding. Weak geometry information is captured by spatial binning using 1×1, 3×1 and 2×2 grids. Readers are referred to [12] for more details.

Saliency map pyramid. We use a state-of-the-art salient object detection model [58] to compute a saliency map for an image, and compute a spatial pyramid of a 8×8 and a 16×16 grid. Each grid cell contains the average saliency value within it. The cells of the spatial pyramid are then concatenated into a 320-D vector.

CNN feature. We use Caffe [28] for fine-tuning the CNN model pre-trained on ImageNet [44]. Images are resized to 256×256 regardless of of their original aspect ratios. The top-left, top-right, bottom-left and bottom-right 227×227 crops of a image are used to augment the training data. We use Caffe's default setting for training the CNN model of [30], but reduce the starting learning rate to 0.001 as in [22]. We stop tuning after around 30 epochs, as the training loss no longer decreases.

4.2. Experimental Settings

For training and testing, we randomly split the SOS dataset into a training set of 5520 images (80% of the SOS dataset) and a testing set of 1380 images. We train linear SVM classifiers for GIST, HOG, IFV and the saliency map pyramid feature (SalPyr). The hyper-parameters of the SVM are determined via 5-fold validation. According to the validation results, we reduce the dimension of GIST, HOG features to 100-D by PCA (Principal Component Analysis), which slightly improves the validation accuracy. For the CNN feature, we directly use the probabilistic output from the fc_8 layer as the final prediction scores.

In addition, we evaluate the performance of the pre-trained CNN without fine-tuning (CNN_wo_FT). We apply the same SVM training procedure on the 4096-D feature output from the fc_7 layer. Moreover, to see how well counting can perform, we evaluate another baseline method that counts the connected components of a binarized saliency map. We use the state-of-the-art salient detection method of [58], and binarize its saliency maps using Otsu's method [39]. Components with areas smaller than $1/100$ of the area of the largest component are removed to suppress small components that can be caused by a cluttered background. Since this counting-based method cannot determine the existence of salient objects in images, we only report its performance in predicting the number of salient objects.

4.3. Results and Analysis

In Table 3, we show the Average Precision (AP) scores of each baseline method. We use the implementation of [19] to calculate AP. The baseline Chance generates random confidence scores for each category, and we report the average AP scores over 100 random trials.

The fine-tuned CNN achieves consistently better performance vs. other baselines over all categories, giving a mean AP score of 0.69. Fine-tuning gives about 15% relative performance gain over the pre-trained CNN feature (CNN_wo_FT). The fine-tuned CNN attains over 90% AP scores in predicting images with no salient object and with a single salient object. We have also tried combining the other features with CNN, but the performance does not further improve.

The IFV feature significantly outperforms the other non-

Table 3: Average Precision (AP) of compared methods. The best scores are shown in red color. The CNN feature significantly outperforms the other baselines.

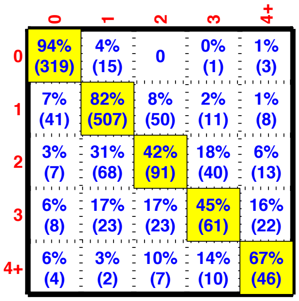
	0	1	2	3	4+	mean
Chance	.28	.48	.19	.12	.07	.23
SalCount	-	.55	.21	.16	.11	-
SalPyr	.41	.62	.36	.21	.09	.34
HOG	.65	.62	.32	.29	.14	.40
GIST	.69	.66	.32	.23	.22	.42
IFV	.84	.69	.32	.24	.44	.50
CNN_wo_FT	.92	.82	.34	.31	.56	.59
CNN	.93	.90	.51	.48	.65	.69

CNN baselines in predicting images with no object and with 4+ objects. The GIST and HOG features give similar performance, and they are consistently better than chance in all categories. The saliency map pyramid (SalPyr) feature achieves 0.36 AP in predicting the images that have two salient objects, outperforming all the other baselines except CNN. However, SalPyr is not as effective as HOG and GIST in predicting the existence of salient objects in an image.

It is not surprising that the performance of the counting-based method is barely better than chance. Counting based on pixel connectivity is only reliable in idealistic cases, where salient objects are separated and well detected. Many images in the SOS dataset have cluttered backgrounds and overlapping foreground objects, making the prediction of the number salient objects a non-trivial task.

Fig. 7, shows the confusion matrix for our best baseline method, using the fine-tuned CNN model. The percentage reported in each cell represents the proportion of images of category A (row number) classified as category B (column number). The accuracy (recall) of category 0 is 94%, matching the human accuracy for this category in our human subitizing test (see Fig. 5). For the remaining categories, there is still a considerable gap between human and machine performance, especially for categories with more than one salient object (compare Fig. 5 with Fig. 7). The recognition accuracy of category 2 is the lowest, and about 49% of the images in this category are confused with neighboring numbers. The performance increase for class 4+ compared to 2-3 may be related to the fact that Images of class 4+ tend to contain large groups of objects. This can make the 4+ images more visually distinctive from images with a much smaller number of objects. Sample results are displayed in Fig. 8.

To gain further insight into the model learned by the best baseline, we used the method of [47] to visualize the fine-tuned CNN classifiers. Sample visualizations are included in the supplementary material. The visualization results indicate that the CNN captures some common visual patterns for each category, especially for categories 2, 3 and 4+.



	0	1	2	3	4+
0	94% (319)	4% (15)	0	0% (1)	1% (3)
1	7% (41)	82% (507)	8% (50)	2% (11)	1% (8)
2	3% (7)	31% (68)	42% (91)	18% (40)	6% (13)
3	6% (8)	17% (23)	17% (23)	45% (61)	16% (22)
4+	6% (4)	3% (2)	10% (7)	14% (10)	67% (46)

Figure 7: Confusion matrix of our method using the fine-tuned CNN feature. Each row corresponds to a ground-truth category. The percentage reported in each cell is the proportion of images of the category A (row number) labeled as category B (column number).

Regarding the speed, the CNN-based method is capable of processing an image within 3 milliseconds using a modern GPU [28]. Therefore, this technique is quite suitable for many time-critical applications.

5. Application I: Salient Object Detection

In this section, we describe a simple application of the SOS technique for improving the accuracy of salient object detection. Salient object detection aims to automatically localize salient objects in an image. However, most salient object detection methods assume the presence of salient objects in an image; as a consequence, these methods can output unexpected results for images that contain no salient object [55].

This suggests that we can exploit our CNN-based SOS method to identify images that contain no salient objects, as a precomputation for salient object detection methods. For a given image, if our SOS method predicts that the image contains zero salient objects, then we do not apply salient object detection methods on that image. We have found that this simple scheme can significantly improve efficiency and reduce false alarms for salient object detection.

5.1. Experiment on the MSO Dataset

Existing salient object detection benchmark datasets lack images that contain zero salient objects. Moverover, these datasets usually have a majority of images where these is only a single salient object. This makes the evaluation settings of these benchmarks too simplified to simulate realistic scenarios. In light of these drawbacks of past datasets, we assembled a Multi-Salient-Object (MSO) dataset. The MSO dataset has more balanced proportions of images with zero salient objects, one salient object, and multiple salient objects. We believe that this dataset provides a more realistic setting to evaluate salient object detection methods. This dataset will be publicly available.

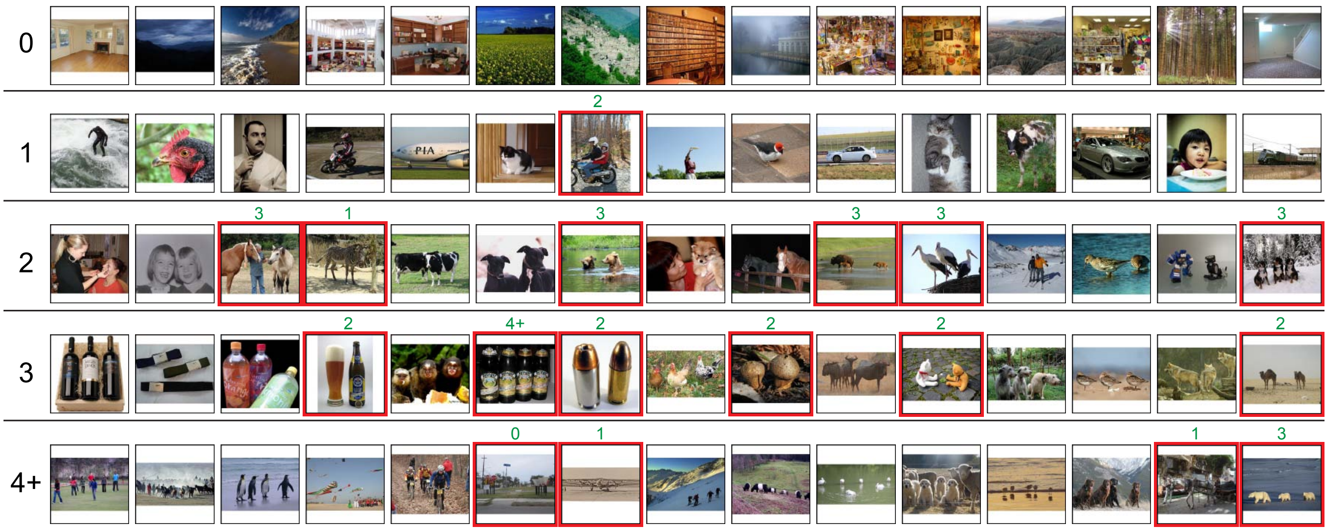


Figure 8: Sample results among the top 100 predictions for each category by the CNN-based subitizing classifier. The images are listed in descending order of confidence. False alarms are shown with red borders and ground-truth labels at the top.

Table 4: Distribution of images with different number salient objects in the MSO dataset.

#Sal. Obj.	0	1	2	3	4+	Total
#Image	338	611	155	100	20	1224

Images of the MSO dataset are taken from the test set of the SOS dataset. We annotated a bounding box for each individual salient object in an image. The number of bounding boxes matches the ground-truth label provided by AMT workers. We removed images with severely overlapping salient objects. We also removed the images for which we find it ambiguous to label the indicated number of salient objects. This leaves us with 1224 images out of 1380 images from our SOS test set. As shown in Table 4, more than 50% images in our MSO dataset contain either zero salient objects or more than one salient objects.

We test two state-of-the-art algorithms on our MSO dataset: Edge Boxes (EB) [59], which is an object proposal method, and LBI [48], which is a salient object detection method. Both methods output a requested number of ranked bounding boxes. We use the Intersection over Union (IOU) to measure the match between predicted bounding boxes and the ground truth. The IOU threshold is set at 0.5, as in the PASCAL challenge [19]. We vary the number of output bounding boxes for each method, and plot Precision-Recall (PR) curves to measure overall performance, as in [48].

Fig. 9 shows the PR curves of the tested models with and without suppressing outputs on the pre-detected background images using our CNN-based subitizing classifier. By employing the subitizing classifier, the PR curve of each method improves, and the Average Precision score attains over 35% relative increase for both of the tested methods. This improvement is attributed to the high accuracy of our CNN-based subitizing classifier in identifying images with no salient object (see Fig. 7). We suspect that further improvement can be obtained by incorporating the subitizing predictions for other categories. For example, for images predicted as containing a single salient object, we can prioritize bounding boxes covering all the salient regions.

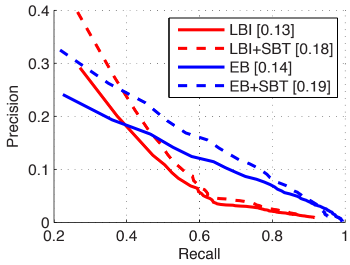


Figure 9: Precison-Recall curves of the compared methods on the MSO dataset. LBI+SBT (EB+SBT *resp.*) denotes the result of LBI (EB *resp.*), with output suppressed for predicted background images using our subitizing classifier. The subitizing classifier improves the precision for both tested methods. The numbers in the brackets are the Average Precision scores.

5.2. Cross-Dataset Generalization

To test how well the performance of our subitizing classifier generalizes to a different dataset for detecting the presence of salient objects in images, we evaluate it on the web thumbnail image test set proposed in [55]. The test set in

Table 5: Recognition accuracy in predicting the presence of salient objects on the thumbnail image dataset [55]. *We show the 5-fold cross validation accuracy reported for [55]. While our method is trained on the MSO dataset, it generalizes well to this other dataset.

	[55]	Ours
Accuracy	82.8%*	**86.5%**

[55] is composed of 5000 thumbnail images from the Web, and 3000 images sampled from the MSRA-B [35] dataset. 50% of these images contain a single salient object, and the rest contain no salient object. Images for MSRA-B are resized to 130 × 130 to simulate thumbnail images [55].

In Table 5, we report the recognition accuracy of our CNN-based subitizing classifier, in comparison with the 5-fold cross-validation accuracy of the best model reported in [55]. Note that our subitizing classifier is trained on a different dataset (SOS), while the compared model is trained on a subset of the tested dataset via cross validation. Our method outperforms the model of [55], and it can output the prediction in a few milliseconds, without resorting to any salient object detection methods. In contrast, the model of [55] requires computing several saliency maps, which takes over 4 seconds per image as reported in [55].

6. Application II: Object Proposal

The goal of an object proposal method is to propose a compact set of image regions that cover all the objects in a scene [59, 2]. Object proposals can significantly improve the efficiency of object detection. Usually, for object detection, a fixed number of object proposals is used for each image, regardless of any content information for the image. However, for simple images, *e.g.* images with a small number of objects, a few proposals may suffice. Thus, we can further improve the accuracy and efficiency of object proposal methods by dynamically allocating a proper number of proposals based on the image content.

We propose to apply our CNN-based subitizing classifier to identify images with different numbers of dominant objects, and selectively reduce the number of retrieved proposals according to the predicted number of salient objects.

We test this approach on the VOC07[19] test set, using three state-of-the-art object proposal methods, Edge Boxes (EB) [59], MCG [4] and Selective Search (SS) [53]. Note that our SOS dataset (used in training our model) does not include images from the VOC07 test set.

Because in the VOC07 dataset, objects are annotated regardless of whether they are salient or not, images predicted as containing no salient object can often have many background objects annotated. Thus, using our CNN-based subitizing classifier, we retrieve $\frac{1}{2}N$ proposals for images identified as containing 1, 2 or 3 salient objects, and N proposals otherwise. We vary the retrieval number N, and calculate the recall rate against the average number of proposals per image.

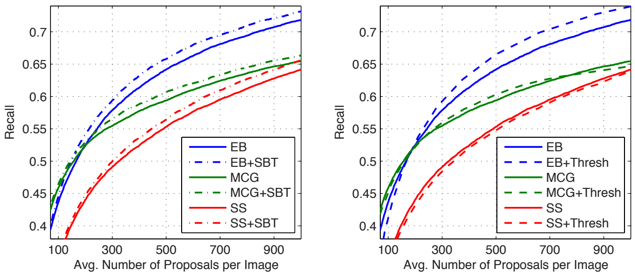


Figure 10: Recall rate against the avg. number of proposals per image on the VOC07 test set, with IOU threshold at 0.7. *Left*: using our subitizing classifier ([method]+SBT) consistently improves the performance of all the tested methods. *Right*: the effect of the naïve global threshold baseline ([method]+Thresh) is more method-dependent.

We compare the proposed approach with a naïve global threshold baseline. For an object proposal method, we discard its output bounding boxes whose confidence scores are below a global threshold T. By varying T, we can calculate the aforementioned evaluation metric.

Fig. 10 shows the results. The global threshold baseline improves EB, but fails to benefit the other two, and it even slightly degrades the performance of SS. In contrast, the proposed simple scheme based on subitizing consistently improves all the tested methods, which shows that incorporating the subitizing information is generally beneficial for the object proposal task.

7. Conclusion

In this work, we introduced the Salient Object Subitizing problem, which aims to predict the existence and the number of salient objects in an image using global image features, without resorting to any localization process. We provided a new image dataset for this problem, and showed that for a substantial proportion of images of our dataset, the inter-subject labeling consistency is high. Several global features were benchmarked, and the CNN feature significantly outperformed the other ones. We demonstrated that the simple application of our subitizing technique can improve state-of-the-art methods for salient object detection and object proposal. In addition, we established a Multiple-Salient-Object dataset for evaluating salient object detection methods in a more general setting.

Acknowledgments. This research was supported in part by US NSF grants 0910908 and 1029430, and gifts from Adobe and NVIDIA.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[3] D. Anoraganingrum. Cell segmentation with median filter and mathematical morphology operation. In *International Conference on Image Analysis and Processing*, 1999.

[4] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman. Interactive object counting. In *European Conference on Computer Vision (ECCV)*. 2014.

[6] J. Atkinson, F. W. Campbell, and M. R. Francis. The magic number 4±0: A new look at visual numerosity judgements. *Perception*, 5(3):327–34, 1976.

[7] T. L. Berg and A. C. Berg. Finding iconic images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2009.

[8] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *European Conference on Computer Vision (ECCV)*. 2012.

[9] S. T. Boysen and E. J. Capaldi. *The development of numerical competence: Animal and human models*. Psychology Press, 2014.

[10] A. B. Chan, Z.-S. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[11] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[12] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference (BMVC)*, 2011.

[13] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[14] J. Choi, C. Jung, J. Lee, and C. Kim. Determining the existence of objects in an image and its application to image thumbnailing. *Signal Processing Letters*, 21(8):957–961, 2014.

[15] D. H. Clements. Subitizing: What is it? why teach it? *Teaching children mathematics*, 5:400–405, 1999.

[16] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[17] H. Davis and R. Pérusse. Numerical competence in animals: Definitional issues, current evidence, and a new research agenda. *Behavioral and Brain Sciences*, 11(04):561–579, 1988.

[18] S. Dehaene. *The number sense: How the mind creates mathematics*. Oxford University Press, 2011.

[19] 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.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[21] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[23] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] H. J. Gross. The magical number four: A biological, historical and mythological enigma. *Communicative & integrative biology*, 5(1):1–2, 2012.

[25] H. J. Gross, M. Pahl, A. Si, H. Zhu, J. Tautz, and S. Zhang. Number-based visual generalisation in the honeybee. *PloS one*, 4(1):e4263, 2009.

[26] B. R. Jansen, A. D. Hofman, M. Straatemeier, B. M. Bers, M. E. Raijmakers, and H. L. Maas. The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology*, 32(2):178–194, 2014.

[27] W. S. Jevons. The power of numerical discrimination. *Nature*, 3:367–367, 1871.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014.

[29] E. L. Kaufman, M. Lord, T. Reese, and J. Volkmann. The discrimination of visual number. *The American journal of psychology*, pages 498–525, 1949.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[31] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[32] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[33] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[35] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[36] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[37] G. Mandler and B. J. Shebo. Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, 111(1):1, 1982.

[38] S. K. Nath, K. Palaniappan, and F. Bunyak. Cell segmentation using coupled level sets and graph-vertex coloring. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2006.

[39] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[40] M. Pahl, A. Si, and S. Zhang. Numerical cognition in bees and other insects. *Frontiers in psychology*, 4, 2013.

[41] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, 2010.

[42] M. Piazza and S. Dehaene. From number neurons to mental arithmetic: The cognitive neuroscience of number sense. *The cognitive neurosciences, 3rd edition*, pages 865–77, 2004.

[43] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), DeepVision Workshop*, 2014.

[44] 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.

[45] C. Scharfenberger, S. L. Waslander, J. S. Zelek, and D. A. Clausi. Existence detection of objects in images for robot vision using saliency histogram features. In *IEEE International Conference on Computer and Robot Vision (CRV)*, 2013.

[46] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[47] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2014.

[48] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[49] I. Stoianov and M. Zorzi. Emergence of a visual number sense in hierarchical generative models. *Nature neuroscience*, 15(2):194–196, 2012.

[50] V. B. Subburaman, A. Descamps, and C. Carincotte. Counting people in the crowd using a generic head detector. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2012.

[51] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[52] L. M. Trick and Z. W. Pylyshyn. Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological review*, 101(1):80, 1994.

[53] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[54] P. O. Vuilleumier and R. D. Rafal. A systematic study of visual extinction between-and within-field deficits of attention in hemispatial neglect. *Brain*, 123(6):1263–1279, 2000.

[55] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li. Salient object detection for searched web images via global saliency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[56] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[57] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[58] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[59] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, 2014.

[60] W. Y. Zou and J. L. McClelland. Progressive development of the number sense in a deep neural network. In *Annual Conference of the Cognitive Science Society (CogSci)*, 2013.