

# Salient Object Subitizing

Jianming Zhang · Shugao Ma · Mehrnoosh Sameki · Stan Sclaroff ·  
Margrit Betke · Zhe Lin · Xiaohui Shen · Brian Price · Radomír Měch

Received: date / Accepted: date

**Abstract** We study the problem of Salient Object Subitizing, *i.e.* predicting the existence and the number of salient objects in an image using holistic cues. This task is inspired by the ability of people to quickly and accurately identify the number of items within the subitizing range (1-4). To this end, we present a salient object subitizing image dataset of about 14K everyday images which are annotated using an online crowdsourcing marketplace. We show that using an end-to-end trained Convolutional Neural Network (CNN) model, we achieve prediction accuracy comparable to human performance in identifying images with zero or one salient object. For images with multiple salient objects, our model also provides significantly better than chance performance without requiring any localization process. Moreover, we propose a method to improve the training of the CNN subitizing model by leveraging synthetic images. In experiments, we demonstrate the accuracy and generalizability of our CNN subitizing model and its applications in salient object detection and image retrieval.

**Keywords** Salient object · Subitizing · Deep learning · Convolutional neural network



**Fig. 1** How fast can you tell the number of prominent objects in each of these images? It is easy for people to identify the number of items in the range of 1-4 by a simple glance. This “fast counting” ability is known as *Subitizing*.

## 1 Introduction

How quickly can you tell the number of **salient** objects in each image in Fig. 1?

As early as the 19th century, it was observed that humans can effortlessly identify the number of items in the range of 1-4 by a glance (Jevons, 1871). Since then, this phenomenon, later coined by Kaufman *et al.* as *Subitizing* (Kaufman *et al.*, 1949), has been studied and tested in various experimental settings (Atkinson *et al.*, 1976; Mandler and Shebo, 1982). It is shown that identifying small numbers up to three or four is highly accurate, quick and confident, while beyond this subitizing range, this sense 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 (Dehaene, 2011; Gross *et al.*, 2009; Davis and Pérusse, 1988; Pahl *et al.*, 2013). This suggests that subitizing may be an inborn numeric capacity of humans and animals. It is speculated that subitizing is a preattentive and parallel process (Dehaene, 2011; Trick and Pylyshyn, 1994; Vuilleumier and Rafal, 2000), and that it can help humans and animals make prompt decisions in basic tasks like navigation, searching and choice making (Piazza and Dehaene, 2004; Gross, 2012).

Jianming Zhang, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke

Computer Science Dept., Boston Univ., Boston, MA USA  
E-mail: {jmzhang,sameki,sclaroff,betke}@bu.edu

Shugao Ma

Oculus Research Pittsburgh, Pittsburgh, PA USA  
E-mail: shugao.ma@oculus.com

Zhe Lin, Xiaohui Shen, Brian Price, Radomír Měch

Adobe Research, San Jose, CA USA  
E-mail: {zlin,xshen,bprice,rmech}@adobe.com



**Fig. 2** Sample images of the proposed SOS dataset. We collected about 14K everyday images, and use Amazon Mechanical Turk (AMT) to annotate the number of salient object of each image. The consolidated annotation is shown on the top of each image group. These images cover a wide range of content and object categories.

Inspired by the subitizing phenomenon, we propose to study the problem of *Salient Object Subitizing* (SOS), *i.e.* predicting the existence and the number (1, 2, 3, and 4+) of salient objects in an image without using any localization process. Solving the SOS problem can benefit many computer vision tasks and applications.

Knowing the existence and the number of salient objects without the expensive detection process 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 machine vision system suppress the object recognition process, until the existence of salient objects is detected, and it can also provide cues for generating a proper number of salient object detection windows for subsequent processing. Furthermore, differentiating between scenes with zero, a single and multiple salient objects can also facilitate applications like image retrieval, iconic image detection (Berg and Berg, 2009), image thumbnailing (Choi et al., 2014), robot vision (Scharfenberger et al., 2013), egocentric video summarization (Lee et al., 2012), snap point prediction (Xiong and Grauman, 2014), *etc.*

In our preliminary work (Zhang et al., 2015a), we presented the first formulation of SOS and an SOS image dataset of about 7K images. The number of salient objects in each image was annotated by Amazon Mechanical Turk (AMT) workers. The resulting annotations from the AMT workers were analyzed in a more controlled offline setting; this analysis showed a high inter-subject consistency in subitizing salient objects in the collected images. In this paper, we follow the same data collection procedure and expand our SOS dataset by approximately doubling the dataset size. This allows us to train more generalizable SOS models and have more robust evaluations. In Fig. 2, we show some sample images in the SOS dataset with the collected groundtruth labels.

We formulate the SOS problem as an image classification task, and aim to develop a method to quickly and accurately predict the existence and the number of generic salient objects in everyday images. We propose to use an end-to-end trained Convolutional Neural Network (CNN) model for our task, and show that an implementation of our method achieves very promising performance. In particular, the CNN-based subitizing model can approach human performance in identifying images with no salient object and with a single salient object. We visualize the learned CNN features and show that these features are quite generic and discriminative for the class-agnostic task of subitizing. Moreover, we empirically validate the generalizability of the CNN subitizing model to unseen object categories.

To further improve the training of the CNN SOS model, we experiment with the usage of synthetic images. We generate a total of 20K synthetic images that contain different numbers of dominant objects using segmented objects and background images. We show that model pre-training using these synthetic images results in an absolute increase of more than 2% in Average Precision (AP) in identifying images with 2, 3 and 4+ salient objects respectively. In particular, for images with 3 salient objects, our CNN model attains an absolute increase of about 6% in AP.

We demonstrate the application of our SOS method in salient object detection and image retrieval. For salient object detection, our SOS model can effectively suppress false object detections on background images and estimate a proper number of detections. By leveraging the SOS model, we attain an absolute increase of about 4% in F-measure over the state-of-the-art performance in unconstrained salient detection (Zhang et al., 2016). For image retrieval, we show that the SOS method can be used to handle queries with object number constraints.

In summary, the key contributions of this work are:

1. We formulate the Salient Object Subitizing (SOS) problem, which aims to predict the number of salient objects in an image without resorting to any object localization process.
2. We provide a large-scale image dataset for studying the SOS problem and benchmarking SOS models.
3. We present a CNN-based method for SOS, and propose to use synthetic images to improve the learned CNN model.
4. We demonstrate applications of the SOS method in salient object detection and image retrieval.

Compared with our preliminary work on SOS (Zhang et al., 2015a), we make several major improvements in this paper: 1) we expand the SOS dataset by doubling the number of images; 2) we attain significantly better performance by leveraging a more advanced CNN architecture, additional real training data and a large number of synthetic training data; 3) we conduct extensive experimental analyses to compare CNN model architectures, visualize the learned CNN features, and validate the generalizability of the SOS model for unseen object categories; 4) in addition to salient object detection, we demonstrate the application of SOS in image retrieval.

## 2 Related Work

**Salient object detection.** Salient object detection aims at detecting dominant objects in a scene. Given a test image, some methods (Achanta et al., 2009; Cheng et al., 2011; Shen and Wu, 2012; Zhang et al., 2015b) generate a saliency map that highlights the overall region of salient objects; other methods (Liu et al., 2011; Gopalakrishnan et al., 2009; Feng et al., 2011; Siva et al., 2013) produce bounding boxes for localization. Ideally, if a salient object detection method can well localize each salient object, then the number of objects can be simply inferred by counting the detection windows. However, many existing salient object detection methods assume the existence of salient objects, and they are mainly tested and optimized for images that contain a single dominant object (Li et al., 2014; Borji et al., 2012). Therefore, salient object detection methods often generate undesirable results on background images, and are prone to fail on images with multiple objects and complex background. Recently, Zhang et al. (2016) proposed a salient object detection method for unconstrained images. Although this method can handle complex images to some extent, we will show that the counting-by-detection approach is less effective than our subitizing method in predicting the number of salient objects.

### Detecting the existence of salient objects.

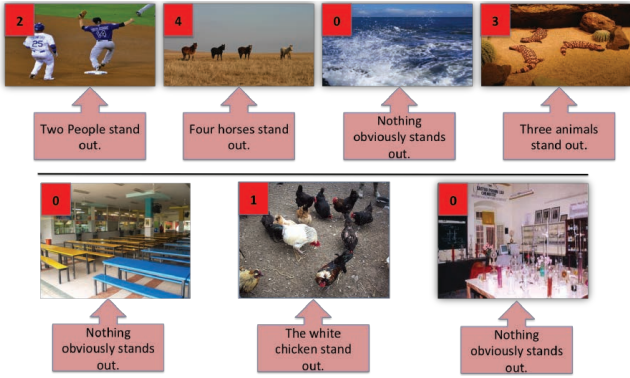
Only a few works address the problem of detecting the existence of salient objects in an image. Wang et al. (2012) use a global feature based on several saliency maps to determine the existence of salient objects in thumbnail images. Their method assumes that an image either contains a single salient object or none. Scharfenberger et al. (2013) use saliency histogram features to detect the existence of interesting objects for robot vision. It is worth noting that the testing images handled by the methods of Wang et al. (2012) and Scharfenberger et al. (2013) are substantially simplified compared to ours, and these methods cannot predict the number of salient objects.

**Automated object counting.** There is a large body of literature about automated object counting based on density estimation (Lempitsky and Zisserman, 2010; Arteta et al., 2014), object detection/segmentation (Subburaman et al., 2012; Nath et al., 2006; Anoraganingrum, 1999) and regression (Chan et al., 2008; Chan and Vasconcelos, 2009). While automated object counting methods are often designed for crowded scenes with many objects to count, the SOS problem aims to discriminate between images with 0, 1, 2, 3 and 4+ dominant objects. Moreover, automated object counting usually focuses on a specific object category (*e.g.* people and cells), and assumes that the target objects have similar appearances and sizes in the testing scenario. On the contrary, the SOS problem addresses category-independent inference of the number of salient objects. The appearance and size of salient objects can vary dramatically from category to category, and from image to image, which poses a very different challenge than the traditional object counting problem.

**Modeling visual numerosity.** Some researchers exploit deep neural network models to analyze the emergence of visual numerosity in human and animals (Stoianov and Zorzi, 2012; Zou and McClelland, 2013). 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.

## 3 The SOS Dataset

We present the Salient Object Subitizing (SOS) dataset, which contains about 14K everyday images. This dataset expands the dataset of about 7K images reported in our preliminary work (Zhang et al., 2015a).



**Fig. 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.

We first describe the collection of this dataset, and then provide a human labeling consistency analysis for the collected images. The dataset is available on our project website<sup>1</sup>.

### 3.1 Image Source

To collect a dataset of images with different numbers of salient objects, we gathered an initial set of images from four popular image datasets, COCO (Lin et al., 2014), ImageNet (Russakovsky et al., 2015), VOC07 (Everingham et al., 2007), and SUN (Xiao et al., 2010). Among these datasets, COCO, ImageNet and VOC07 are designed for object detection, while SUN is for scene classification. Images from COCO and VOC07 often have complex backgrounds, but their content is limited to common objects and scenes. ImageNet contains a more diverse set of object categories, but most of its images have centered dominant objects with relatively simpler backgrounds. In the SUN dataset, many images are rather cluttered and do not contain any salient objects. We believe that combining images from different datasets can mitigate the potential data bias of each individual dataset.

This preliminary set is composed of about 30000 images in total. There are about 5000 images from SUN, 5000 images from VOC07 respectively, 10000 images are from COCO and 10000 images from ImageNet. For VOC07, the whole training and validation sets are included. We limited the number of images from the SUN dataset to 5000, 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 background images. The 5000 images were randomly sampled from SUN. For the COCO and Im-

**Table 1** Distribution of images in the SOS dataset

category	COCO	VOC07	ImageNet	SUN	total
0	616	311	371	1963	3261
1	2504	1691	1516	330	6041
2	585	434	935	76	2030
3	244	106	916	43	1309
4+	371	182	475	38	1066
total	4320	2724	4213	2450	13707



**Fig. 4** Sample images with divergent labels. These images are a bit ambiguous about what should be counted as an individual salient object. We exclude this type of images from the final SOS dataset.

geNet datasets<sup>2</sup>, we used the bounding box annotations to split the dataset into four categories for 1, 2, 3 and 4+, and then sampled 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 ambiguous cases for counting, *e.g.* counting a man riding a horse as one or two objects. We expected that ambiguous images would lead to divergent annotations.

Each task, or HIT (Human Intelligence Task) was composed of five to ten images with a two-minute time limit, and the compensation was one to two cents per task. All the images in one task were displayed at the same time. The average completion time per image was about 4s. We collected five annotations per image from distinct workers. About 800 workers contributed to this dataset. The overall cost for collecting the annotation is about 600 US dollars including the fees paid to the AMT platform.

<sup>1</sup> <http://www.cs.bu.edu/groups/ivc/Subitizing/>

<sup>2</sup> 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)

**Fig. 5** Averaged confusion matrix of our offline human subitizing test. Each row corresponds to a groundtruth 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.

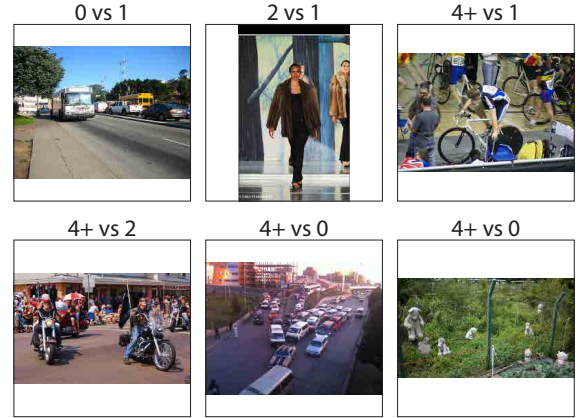
**Table 2** Human subitizing accuracy in matching category labels from Mechanical Turk workers.

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

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 about 14K 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. As expected, the majority of the images from the SUN dataset belong to the “0” category. The ImageNet dataset contains significantly more images with two and 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 minutes 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 important 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 dataset.



**Fig. 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).

To answer these questions, we conducted a more controlled offline experiment based on common experimental settings in the subitizing literature (Atkinson et al., 1976; Mandler and Shebo, 1982). 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 78% agreement, and about 13% of images in this category are classified as category 0.

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 by Convolutional Neural Network

Subitizing is believed to be a holistic sense of the number of objects in a visual scene. This visual sense can discriminate between the visual patterns possessed by different numbers of objects in an image (Jansen et al., 2014; Mandler and Shebo, 1982; Clements, 1999; Boyesen and Capaldi, 2014). This inspires us to propose a learning-based discriminative approach to address the SOS problem, without resorting to any object localization or counting process. In other words, we aim to train image classifiers to predict the number of salient objects in a image.

Encouraged by the remarkable progress made by the CNN models in computer vision (Girshick et al., 2014; Krizhevsky et al., 2012; Razavian et al., 2014; Sermanet et al., 2014), we use the CNN-based method for our problem. Girshick et al. (2014) suggest that given limited annotated data, fine-tuning a pre-trained CNN model can be an effective and highly practical approach for many problems. Thus, we adopt fine-tuning to train the CNN SOS model.

We use the GoogleNet architecture (Szegedy et al., 2015), which has significantly fewer parameters than

the AlexNet model in our previous SOS paper (Zhang et al., 2015a). However, GoogleNet is shown to substantially outperform AlexNet in image classification tasks and it also compares favorably with the widely used the VGG16 (Simonyan and Zisserman, 2015) architecture in terms of speed and classification accuracy. We fine-tune the GoogleNet model pre-trained on ImageNet (Russakovsky et al., 2015) using Caffe (Jia et al., 2014). The output layer of the pre-trained GoogleNet model is replaced by a fully connected layer which outputs a 5-D score vector for the five categories: 0, 1, 2, 3 and 4+. We use the Softmax loss and the SGD solver of Caffe to fine-tune all the parameters in the model. More training details are provided in Sec. 5.

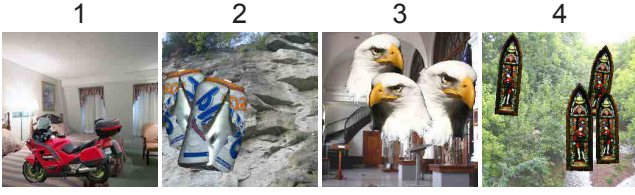
### 4.1 Leveraging Synthetic Images for CNN Training

Collecting and annotating real image data is a rather expensive process. Moreover, the collected data may not have a balanced distribution over all the categories. In our SOS dataset, over 2/3 images belong to the “0” or “1” category. For categories with insufficient data, the CNN model training may suffer from overfitting and lead to degraded generalizability of the CNN model.

Leveraging synthetic data can be a economical way to alleviate the burden of image collection and annotation (Stark et al., 2010; Sun and Saenko, 2014; Jaderberg et al., 2014). In particular, some recent works (Jaderberg et al., 2014; Peng et al., 2015) successfully exploit synthetic images to train modern CNN models for image recognition tasks. While previous works focus on generating realistic synthetic images (*e.g.* using 3D rendering techniques (Peng et al., 2015)) to train CNN models with zero or few real images data, our goal is to use synthetic images as an auxiliary source to improve the generalizability of the learned CNN model.

We adopt a convenient *cut-and-past* approach to generate synthetic SOS image data. Given a number  $N$  in the range of 1-4, a synthetic image is generated by pasting  $N$  cutout objects on a background scene image. Cutout objects can be easily obtained from existing image datasets with segmentation annotations or image sources with isolated object photos (*e.g.* stock image databases). In this work, we use the public available salient object dataset THUS10000 (Cheng et al., 2015) for generating cutout objects and the SUN dataset (Xiao et al., 2010) as the source for background images. The THUS10000 dataset covers a wide range of object categories so that we can obtain sufficient variations in the shape and appearance of foreground objects.

In THUS10000, an image may contain multiple salient objects and some of them are covered by a



**Fig. 7** Sample synthetic images with the given numbers of salient objects on the top. Although the synthetic images look rather unrealistic, they are quite visually consistent with the given numbers of salient objects. By pre-training the CNN SOS model on these synthetic images, we expect that the CNN model can better learn the intra-class variations in object category, background scene type, object position and inter-object occlusion.

single segmentation mask. To generate consistent synthetic SOS image data, we automatically filter out this type of images using the CNN SOS model trained on real data. To do this, we remove the images whose confidence scores for containing one salient object are less than 0.95. Similarly, we filter out the images with salient objects from the SUN dataset, using a score threshold of 0.95 for containing no salient object.

When generating a synthetic image, we randomly choose a background image and resize it to  $256 \times 256$  regardless of its original aspect ratio. Then, we pick a cutout object and generate a reference object by resizing it to a randomly generated scale relative to 256 based on the largest dimension of the object. The reference scale is uniformly sampled in the range  $[0.4, 0.8]$ . After that, we apply random horizontal flipping and mild geometric transforms (scaling and rotation) on the reference object each time we past a copy of it to a random position on the background image. Mild scalings are uniformly sampled in the range  $[0.85, 1.15]$  and mild rotations are uniformly sampled in the angular range  $[-10, 10]$  degrees. The synthetic image contains  $N$  ( $N \in [1, 4]$ ) copies of the same cutout object. Pasting different cutout objects together is empirically found inferior to our method, probably because some cutout objects may appear more salient than the other ones when they are put together, resulting in images that visually inconsistent with the given number. Finally, we reject this image if any of the pasted objects is occluded by more 50% of its area.

Example synthetic images are shown in Fig. 7. Our synthetic images look rather unrealistic, since we do not consider any contextual constraints between scene types and object categories. However, for the SOS task, these images often look quite consistent with the given numbers of salient objects. We expect that our CNN model should learn generic features for SOS irrespective of semantics of the visual scenes. Thus, these synthetic images may provide useful intra-class variations

in object category, background scene type, as well as object position and inter-object occlusion.

To leverage the synthetic images, we fine-tune the CNN model on the synthetic data before fine-tuning on the real data. The two-stage fine-tuning scheme can be regarded as a domain adaptation process, which transfers the learned features from the synthetic data domain to the real data domain. Compared with combining the real and synthetic images into one training set, we find that our two-stage fine-tuning scheme works significantly better (see Sec. 5).

## 5 Experiments

### 5.1 Experimental Setting

For training and testing, we randomly split the SOS dataset into a training set of 10,966 images (80% of the SOS dataset) and a testing set of 2741 images.

**CNN model training details.** For fine-tuning the GoogleNet CNN model, images are resized to  $256 \times 256$  regardless of their original aspect ratios. Standard data augmentation methods like horizontal flipping and cropping are used. We set the batch size to 32 and fine-tune the model for 8000 iterations. The fine-tuning starts with a learning rate of 0.001 and we multiply it by 0.1 every 2000 iterations. At test time, images are resized to  $224 \times 224$  and the output softmax scores are used for evaluation.

For pre-training using the synthetic images, we generate 5000 synthetic images for each number in 1-4. Further increasing the number of synthetic images does not increase the performance. We also include the real background images (category “0”) in the pre-training stage. The same model training setting is used as described above. When fine-tuning using the real data, we do not reset the parameters of the top fully-connected layer, because we empirically find that it otherwise leads to slightly worse performance.

**Compared methods.** We evaluate our method and several baselines as follows.

- CNN\_Syn\_FT: The full model fine-tuned using the two-stage fine-tuning scheme with the real and synthetic image data.
- CNN\_Syn\_Aug: The model fine-tuned on the union of the synthetic and the real data. This baseline corresponds to the data augmentation scheme in contrast to the two-stage fine-tuning scheme for leveraging the synthetic image data. This baseline is to validate our two-stage fine-tuning scheme.
- CNN\_FT: The CNN model fine-tuned on the real image data only.

**Table 3** Average Precision (%) of compared methods. The best scores are shown in bold. The training and the testing are repeated for five times for all CNN-based methods, and mean and std of the AP scores are reported.

	0	1	2	3	4+	mean
Chance	27.5	46.5	18.6	11.7	9.7	22.8
SalPyr	46.1	65.4	32.6	15.0	10.7	34.0
HOG	68.5	62.2	34.0	22.8	19.7	41.4
GIST	67.4	65.0	32.3	17.5	24.7	41.4
SIFT+IVF	83.0	68.1	35.1	26.6	38.1	50.1
CNN_woFT	92.2±0.2	84.4±0.2	40.8±1.9	34.1±2.7	55.2±0.6	61.3±0.2
CNN_FT	<b>93.6±0.3</b>	<b>93.8±0.1</b>	75.2±0.2	58.6±0.8	71.6±0.5	78.6±0.2
CNN_Syn	79.2±0.5	85.6±0.2	37.4±0.8	34.8±2.6	33.0±1.1	54.0±0.6
CNN_Syn_Aug	92.1±0.4	92.9±0.1	75.0±0.4	58.9±0.6	69.8±0.8	77.8±0.3
CNN_Syn_FT	<b>93.5±0.1</b>	<b>93.8±0.2</b>	<b>77.4±0.3</b>	<b>64.3±0.2</b>	<b>73.0±0.5</b>	<b>80.4±0.2</b>

- CNN\_Syn: The CNN model fine-tuned on the synthetic images only. This baseline reflects how close the synthetic images are to the real data.
- CNN\_wo\_FT: The features of the pre-trained GoogleNet without fine-tuning. For this baseline, we fix the parameters of all the hidden layers during fine-tuning. In other words, only the output layer is fine-tuned.

Furthermore, we benchmark several commonly used image feature representations for baseline comparison. For each feature representation, we train a one-vs-all multi-class linear SVM classifier on the training set. The hyper-parameters of the SVM are determined via five-fold cross-validation.

- GIST. The GIST descriptor (Torralba et al., 2003) is computed based on 32 Gabor-like filters with varying scales and orientations. We use the implementation by Torralba et al. (2003) to extract a 512-D GIST feature, which is a concatenation of averaged filter responses over a  $4 \times 4$  grid.
- HOG. We use the implementation by Felzenszwalb et al. (2010) to compute HOG features. Images are first resized to  $128 \times 128$ , and HOG descriptors are computed on a  $16 \times 16$  grid, with the cell size being  $8 \times 8$ . 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.
- SIFT with the Improved Fisher Vector Encoding (SIFT+IVF). We use the implementation by Chatfield et al. (2011). The codebook size is 256, and the dimensionality of SIFT descriptors is reduced to 80 by PCA. Hellinger’s kernel and L2-normalization are applied for the encoding. Weak geometry information is captured by spatial binning using  $1 \times 1$ ,  $3 \times 1$  and  $2 \times 2$  grids. To extract dense SIFT, we use the VLFeat Vedaldi and Fulkerson (2008) im-

plementation. Images are resized to  $256 \times 256$ , and a  $8 \times 8$  grid is used to compute a 8192-D dense SIFT feature, with a step size of 32 pixels and a bin size of 8 pixels. Similar to HOG, combining SIFT features of different scales does not improve the performance.

- Saliency map pyramid (SalPyr). We use a state-of-the-art CNN-based salient object detection model (Zhao et al., 2015) to compute a saliency map for an image. Given a saliency map, we construct a spatial pyramid of a  $8 \times 8$  layer and a  $16 \times 16$  layer. Each grid cell represents the average saliency value within it. The cells of the spatial pyramid are then concatenated into a 320-D vector.

**Evaluation metric.** We use average precision (AP) as the evaluation metric. We use the implementation provided in the VOC07 challenge Everingham et al. (2007) to calculate AP. For each the CNN-based method, we repeat the training for five times and report both the mean and the standard deviation (std) of the AP scores. This will give a sense of statistical significance when interpreting the difference between CNN baselines.

## 5.2 Results

The AP scores of different features and CNN baselines are reported in Table 3. The baseline Chance in Table 3 refers to the performance of random guess. To evaluate the random guess baseline, we generate random confidence scores for each category, and report the average AP scores over 100 random trials.

All methods perform significantly better than random guess in all categories. Among manually crafted features, SalPyr gives the worst mean AP (mAP) score, while SIFT+IVF performs the best, outperforming SalPyr by 16 absolute percentage points in mAP. SIFT+IVF is especially more accurate than



	0	1	2	3	4+
0	93% (615)	5% (32)	0% (3)	1% (6)	1% (8)
1	2% (29)	93% (1103)	3% (40)	1% (8)	1% (7)
2	1% (4)	18% (77)	67% (286)	13% (54)	1% (5)
3	3% (8)	6% (16)	15% (38)	63% (157)	13% (32)
4+	7% (15)	3% (7)	2% (5)	25% (53)	62% (133)

(a) SOS

	0	1	2	3	4+
0	62% (410)	31% (208)	6% (40)	1% (5)	0% (1)
1	3% (32)	85% (1006)	12% (139)	1% (9)	0% (1)
2	1% (5)	22% (93)	69% (292)	8% (35)	0% (1)
3	2% (5)	13% (33)	29% (74)	50% (125)	6% (14)
4+	2% (5)	13% (28)	23% (48)	30% (63)	32% (69)

(b) Counting

**Fig. 8** Subitizing *vs.* counting. (a) Confusion matrix of our CNN SOS method CNN\_Syn\_FT. Each row corresponds to a groundtruth category. The percentage reported in each cell is the proportion of images of the category A (row number) labeled as category B (column number). (b) Confusion matrix of counting using the salient object detection method by Zhang et al. (2016).

other non-CNN features in identifying images with 0 and 4+ salient objects.

The CNN feature without fine-tuning (CNN\_wo\_FT) outperforms SIFT+IFV by over 10 absolute percentage points in mAP. Fine-tuning (CNN\_FT) further improves the mAP score by 17 absolute percentage points, leading to a mAP score of 78.6%. CNN\_wo\_FT attains comparable performance to CNN\_FT in identifying background images, while it is significantly worse than CNN\_FT in the other categories. This suggests that the CNN feature trained on ImageNet is good for inferring the presence of salient objects, but not very effective at discriminating images with different numbers of salient objects.

Pre-fine-tuning using the synthetic images (CNN\_Syn\_FT) further boosts the performance of CNN\_FT by about 2 absolute percentage points in mAP. The performance is improved in category “2”, “3” and “4+”, where training images are substantially fewer than categories “0” and “1”. In particular, for category “3” the AP score is increased by about 6 absolute percentage points. The usefulness of the synthetic images may be attributed to the fact they can provide more intra-class variations in object category, scene type and the spatial relationship between objects. This is especially helpful when there is not enough real training data to cover the variations.

Using synthetic images alone (CNN\_Syn) gives reasonable performance, a mAP score of 54.0%. It outperforms SIFT+IVF, the best non-CNN baseline trained on the real data. However, it is still much worse than the CNN model trained on the real data. This gives a sense of the domain shift between the real and the synthetic data. Directly augmenting the training data with the synthetic images does not improve and even slightly

worsens the performance (compare CNN\_Syn\_Aug and CNN\_FT in Table 3). We believe that this is due to the domain shift and our two-stage fine-tuning scheme can better deal with this issue.

Fig. 8 (a) shows the confusion matrix for our best method CNN\_Syn\_FT. 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” and “1” is both about 93%, which is close to the human accuracy for these categories in our human subitizing test (see Fig. 5). For the remaining categories, there is still a considerable gap between human and machine performance. According to Fig. 8 (a), our SOS model tends to make mistakes by misclassifying an image into a nearby category. Sample results are displayed in Fig. 9. Despite the diverse object appearance and image background, our SOS model gives reasonable performance.

### 5.3 Analysis

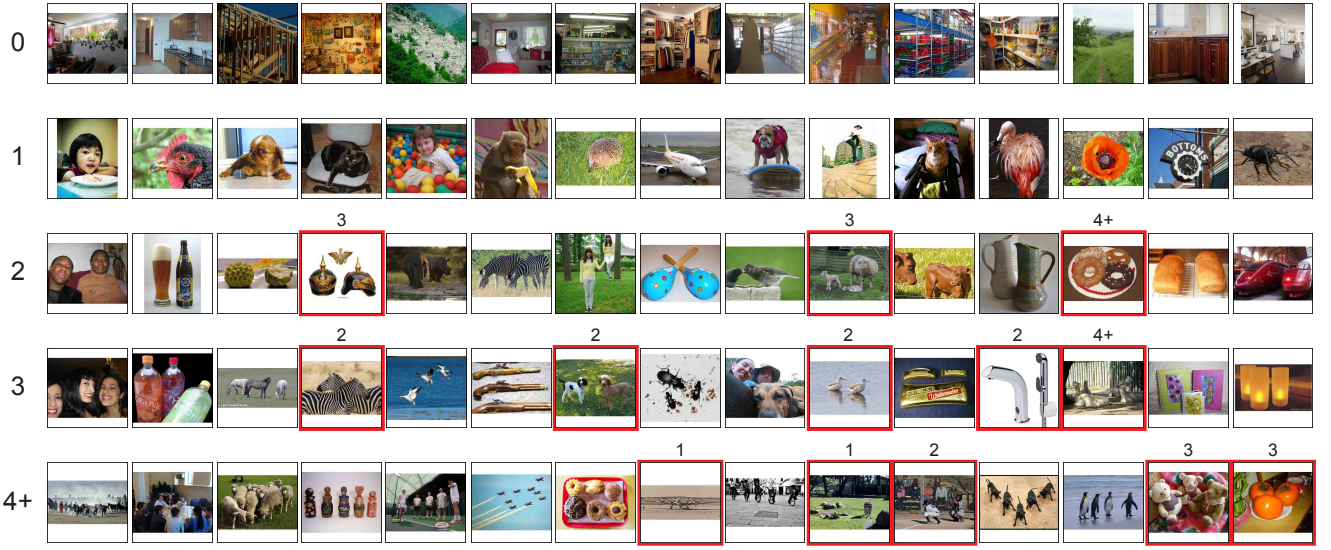
To gain a better understanding of our SOS method, we further investigate the following questions.

#### How does subitizing compare to counting?

Counting is a straightforward way of getting the number of items. To compare our SOS method with a counting-by-detection baseline, we use a state-of-the-art salient object detection method designed for unconstrained images (Zhang et al., 2016). This unconstrained salient object detection method, denoted as USOD, leverages a CNN-based model for bounding proposal generation, followed by a subset optimization method to extract a highly reduced set of detection windows. A parameter of USOD is provided to control the operating point for the precision-recall tradeoff. We pick an operating point that gives the best F-score<sup>3</sup> on the Multi-Salient-Object (MSO) dataset (Zhang et al., 2015a) in this experiment.

The confusion matrix of the counting baseline is shown in Fig. 8 (b). Compared with the SOS method (see Fig. 8 (a)), the counting baseline performs significantly worse in all categories except “2”. In particular, for “0” and “4+”, the counting baseline is worse than the SOS method by about 30 absolute percentage points. This indicates that for the purpose of number prediction, the counting-by-detection approach can be a suboptimal option. We conclude that there are at least two reasons for this outcome. First, it is difficult to pick a fixed score threshold (or other equivalent parameters) of an object detection system that works best for all

<sup>3</sup> The F-score is computed as  $\frac{2RP}{(R+P)}$ , where  $R$  and  $P$  denote recall and precision respectively.



**Fig. 9** Sample results among the top 100 predictions for each category by our CNN SOS method CNN\_Syn\_FT. The images are listed in descending order of confidence. False alarms are shown with red borders and groundtruth labels at the top.

**Table 4** Mean average precision (%) scores for different CNN architectures. Training and test are run for five times and the mean and the std of mAP scores are reported.

	AlexNet	VGG16	GoogleNet
w/o Syn. Data	70.1 $\pm$ 0.2	77.5 $\pm$ 0.3	78.6 $\pm$ 0.2
with Syn. Data	71.6 $\pm$ 0.5	80.2 $\pm$ 0.3	80.4 $\pm$ 0.3

images. Even when an object detector gives a perfect ranking of window proposals for each image, the scores may not be well calibrated across different images. Second, the post-processing step for extracting detection results (*e.g.* non-maximum suppression) is based on the idea of suppressing severely overlapping windows. However, this spatial prior about detection windows can be problematic when significant inter-object occlusion occurs. In contrast, our SOS method bypass the detection process and discriminates between different numbers of salient objects based on holistic cues.

**How does the CNN model architecture affect the performance?** Besides GoogleNet, we evaluate another two popular architectures, AlexNet (Krizhevsky et al., 2012) and VGG16 (Simonyan and Zisserman, 2015). The mAP scores with and without using synthetic images are summarized in Table 4 for each architecture. VGG16 and GoogleNet have very similar performance, while AlexNet performs significantly worse. Pre-training using synthetic images has a positive effect on all these architectures, indicating that it is generally beneficial to leverage synthetic images for this task. The baseline of AlexNet without synthetic image can be regarded as the best model reported by Zhang et al. (2015a). In this sense, our

**Table 5** The effect of using the synthetic images when different numbers of real data are used in CNN training. For each row, the same set of synthetic images are used. Training and test are run for five times and the mean and the std of mAP scores are reported. By using the synthetic images, competitive performance is attained even when the size of the real data is significantly reduced.

	w/o syn.	with syn.
25% real data	71.6 $\pm$ 0.2	76.3 $\pm$ 0.4
50% real data	75.3 $\pm$ 0.3	78.2 $\pm$ 0.4
100% real data	78.6 $\pm$ 0.2	80.4 $\pm$ 0.3

current best method using GoogleNet and synthetic image outperforms the previous best model by 10 absolute percentage points. Note that the training and testing image sets used by Zhang et al. (2015a) are subsets of the training and testing sets of our expanded SOS dataset. Therefore, the scores reported by Zhang et al. (2015a) are not comparable to the scores in this paper<sup>4</sup>.

**Does the usage of synthetic images reduce the need for real data?** To answer this question, we vary the amount of real data used in the training, and report the mAP scores in Table 5. We randomly sample 25% and 50% of the real data for training the model. This process is repeated for five times. When fewer real data are used, the performance of our CNN SOS method declines much slower with the help of the synthetic images. For example, when only 25% real data are used, leveraging the synthetic images can provide an

<sup>4</sup> When evaluated on the test set used by Zhang et al. (2015a), our best method GoogleNet\_Syn\_FT achieves a mAP score of 85.0%

absolute performance gain of about 5% in mAP, leading to a mAP score of 76%. However, without using the synthetic images, doubling the size of the training data (50% real data) only achieves a mAP score of 75%. This suggests that we can achieve competitive performance at a much lower cost at data collection by leveraging the synthetic images.

**What is learned by the CNN model?** By fine-tuning the pre-trained CNN model, we expect that the CNN model will learn discriminative and generalizable feature representations for subitizing. To visualize the new feature representations learned from our SOS data, we first look for features that are substantially distinct from the ones of the original network trained on ImageNet. For GoogleNet, we consider the output layer of the last inception unit (inception\_5b/output), which has 1024 feature channels. For each feature channel of this layer, we use the maximum activation value on an image to rank the images in the SOS test set. We hypothesize that if two feature channels represent similar features, then they should result in similar image rankings. Given the  $i$ -th feature channel of this layer in GoogleNet\_Syn\_FT, we compute the maximum Spearman’s rank correlation coefficient between its image ranking  $R_i$  and the image ranking  $\hat{R}_j$  using the  $j$ -th channel of the original GoogleNet:

$$S_i = \max_{j=1,2,\dots,1024} \rho(R_i, \hat{R}_j), \quad (1)$$

where  $\rho$  denotes Spearman’s rank correlation coefficient, whose range is  $[-1, 1]$ . A low value of  $S_i$  means that the  $i$ -th feature channel of our fine-tuned model gives a very different image ranking than any feature channels from the original CNN model. In our case, none of the values of  $S_i$  is negative. Fig. 10 (a) shows the histogram of  $S_i$ . We choose the feature channels with  $S_i$  values less than 0.3 as the most novel features learned from the SOS data.

After that, we visualize each of the novel feature channels by showing the top nine image patches in our SOS test set that correspond to the highest feature activations for that channel. The spatial resolution of inception\_5b/output is  $7 \times 7$ . For an activation unit on the  $7 \times 7$  map, we display the image patch corresponding to the receptive field of the unit. Since the theoretic receptive field of the unit is too large, we restrict the image patch to be 60% of the size ( $0.6W \times 0.6H$ ) of the whole image.

Fig. 10 (b) shows the visualization results of some of the novel feature representations learned by our CNN SOS model. We find that these newly learned feature representations are not very sensitive to the categories of the objects, but they capture some general visual patterns related to the subitizing task. For example,

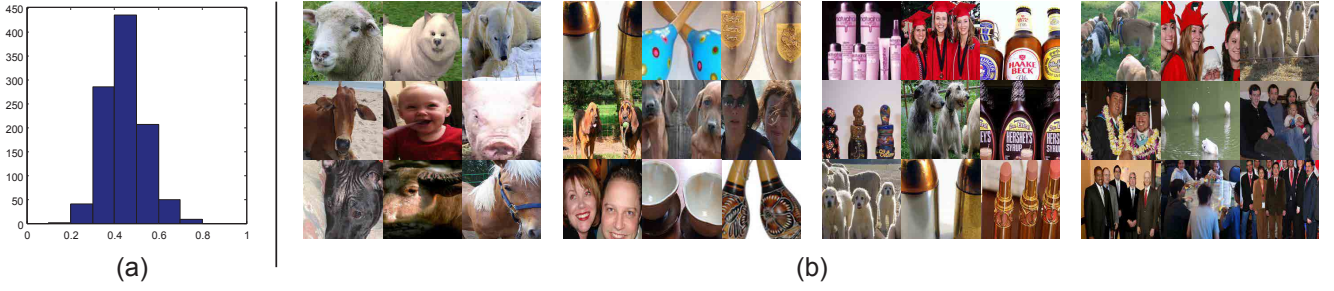
in Fig. 10 (b), the feature corresponding to the first block is about a close-up face of either a person or an animal. Detecting a big face at this scale indicates that the image is likely to contain only a single dominant object. The feature corresponding to the second block is about a pair of objects appearing side by side, which is also a discriminative visual pattern for identifying images with two dominant objects. These visualization results suggest that our CNN model has learned some category-independent and discriminative features for SOS.

**How does the SOS method generalize to unseen object categories?** We would like to further investigate how our CNN SOS model can generalize to unseen object categories. To get category information for the SOS dataset, we ask AMT workers to label the categories of dominant objects for each image in our SOS dataset. We consider five categories: “animal”, “food”, “people”, “vehicle” and “other”. An image may contain multiple labels (*e.g.* an image with an animal and a person). For each image, we collect labels from three different workers and use the majority rule to decide the final labels.

To test the generalizability of our CNN model to unseen object categories, we use the Leave-One-Out (LOO) approach described as follows. Given category  $\mathcal{A}$ , we remove all the images with the label  $\mathcal{A}$  from the original training set, and use them as the testing images. The original test images for “0” are also included. Two other baselines are provided. The first is a chance baseline, which refers to the performance of random guess. We generate random confidence scores for each category, and report the average AP scores over 100 random trials. Note that we have class imbalance in the test images, so the AP scores of random guess tend to be higher for categories with more images. The second baseline reflects the performance for category  $\mathcal{A}$  when full supervision is available. We use five-fold cross-validation to evaluate this baseline. In each fold, 1/5 of the images with the label  $\mathcal{A}$  are used for testing, and all the remaining images are used for training. The average AP scores are reported. In this experiment, we do not use the synthetic images because they do not have category labels.

The results are reported in Table 6. For each category, the CNN model trained without that category (CNN-LOO) gives significantly better performance than the Chance baseline. This validates that the CNN model can learn category-independent features for SOS and it can generalize to unseen object categories to some extent. Training with full supervision (CNN-Full) further improves over CNN-LOO by a substantial margin, which indicates that it is still important to





**Fig. 10** Feature visualization of the inception\_5b/output layer in our GoogleNet\_Syn\_FT model. We aim to visualize the new feature representations learned from our SOS data. (a) shows the histogram of  $S_i$ , which measures how distinct a feature channel of our model is from the feature representations of the original ImageNet model (see text for more details). Lower values of  $S_i$  indicates higher distinctness, and we choose those feature channels with  $S_i < 0.3$  for visualization (b) shows the visualization of some new feature representations learned by our SOS model. Each block displays the top nine image patches in our SOS test set that correspond to the highest feature activations for a novel feature channel. These visualization results suggest that our CNN model has learned some category-independent and discriminative features for SOS. For example, the first block corresponds to a feature about a close-up face, and the second block shows a feature of a pair of objects appearing side by side.

**Table 6** Cross-category generalisation test. The CNN-LOO refers to the AP scores (%) on the unseen object category. CNN-Full serves as an upper bound of the performance when the images of that object category are used in the training (see text for more details). The number following each category name is the number of images with that category label.

		0	1	2	3	4+	mean
animal (4101)	Chance	16.6	53.6	21.1	12.6	8.8	22.5
	CNN-LOO	89.3±0.2	87.2±0.3	42.8±1.0	36.9±2.6	58.3±1.0	62.9±0.5
	CNN-Full	95.0±1.7	94.8±0.4	72.8±2.0	57.9±2.8	71.8±4.0	78.5±1.3
food (372)	Chance	67.6	16.9	8.1	13.1	8.2	22.8
	CNN-LOO	95.7±0.2	70.8±1.3	50.3±0.8	56.8±1.3	39.7±1.4	62.7±0.5
	CNN-Full	97.7±0.4	85.9±7.2	61.1±11.2	67.8±12.4	62.8±8.3	75.1±4.1
people (3786)	Chance	17.5	50.7	21.7	10.9	13.1	22.8
	CNN-LOO	86.7±0.3	84.9±0.5	47.6±0.5	31.6±1.3	56.7±1.2	61.5±0.5
	CNN-Full	94.4±1.3	94.8±0.7	82.5±1.0	62.8±6.1	83.9±2.8	83.7±1.3
vehicle (1150)	Chance	40.6	56.1	8.3	3.4	4.4	22.6
	CNN-LOO	91.0±0.3	92.2±0.3	42.4±2.2	16.3±0.9	47.4±0.9	57.9±0.4
	CNN-Full	96.1±0.7	96.1±0.7	62.2±9.2	25.6±14.2	55.4±20.6	67.1±6.4
other (1401)	Chance	36.4	35.4	14.8	18.6	11.2	23.3
	CNN-LOO	87.0±0.4	78.0±0.7	56.7±0.4	49.9±0.9	50.2±0.8	64.4±0.4
	CNN-Full	93.4±0.4	90.5±2.5	70.8±7.2	63.0±3.2	60.2±8.3	75.6±2.8

use a training set that covers a diverse set of object categories.

## 6 Applications

### 6.1 Salient Object Detection

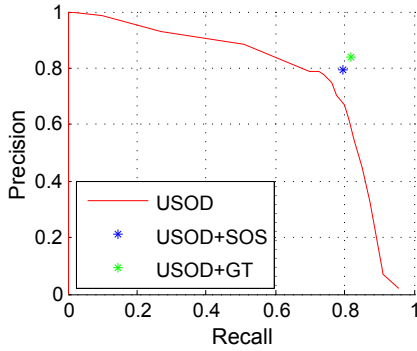
In this section, we demonstrate the usefulness of SOS for unconstrained salient object detection (Zhang et al., 2016). Unconstrained salient object detection aims to detect salient objects in unconstrained images where there can be multiple salient objects or no salient objects. Compared with the constrained setting, where there exists one and only one salient object, the unconstrained setting pose new challenges of handling background images and determining the number of salient

objects. Therefore, SOS can be used to cue a salient object detection method to suppress the detection or output the right number of detection windows for unconstrained images.

Given a salient object detection method, we leverage our CNN SOS model by a straightforward approach. We assume that the salient object detection method provides a parameter (*e.g.* the threshold for the confidence score) for trade-off between precision and recall. We call this parameter as a PR parameter. For an image, we first predict the number of salient objects  $N$  using our CNN SOS model, then we use grid search to find such a value of the PR parameter that no more than  $N$  detection windows are output.

**Dataset.** Most existing salient object detection datasets lack background images or images containing





**Fig. 11** Precision-Recall curve of USOD, and the performance of USOD+SOS and USOD+GT.

multiple salient objects. In this experiment, We use the Multi-Salient-Object (MSO) dataset (Zhang et al., 2015a). The MSO dataset has 1224 images, all of which are from the test set of the SOS dataset, and it has a substantial proportion of images that contain no salient object or multiple salient objects.

**Compared methods.** We test our SOS model on the unconstrained object detection method proposed (denoted as USOD) by Zhang et al. (2016), which achieves state-of-the-art performance on the MSO dataset. The baseline USOD method is composed of a CNN-based object proposal model and a subset optimization formulation for post-processing the bounding box proposals. We use an implementation provided by Zhang et al. (2016), which uses the GoogleNet architecture for proposal generation. The USOD method provides a PR parameter to control the number of detection windows. We use the predicted number by our SOS model to cue USOD, and denote this method as USOD+SOS. We also use the groundtruth number to show the upper-bound of the performance gain using subitizing, and denote this baseline as USOD+GT.

**Evaluation metrics.** We report the precision, the recall and the F-measure. The F-measure is calculated as  $2 \frac{PR}{P+R}$ , where  $P$  and  $R$  denote the precision and the recall respectively. For the baseline USOD method, we tune its PR parameter so that the its F-measure is maximized.

**Results.** The results are reported in Table 7. Fig. 11 shows the PR curve of USOD compared to the precision and recall rates of USOD+SOS and USOD+GT. As we can see, USOD+SOS significantly outperforms the baseline USOD, obtaining an absolute increase of about 4% in F-measure. This validates the benefit of adaptively tuning the PR parameter based on the SOS model. When the groundtruth number of objects is used (USOD+GT), another absolute increase of 3% can be attained, which is the upper bound for the performance improvement. Table 7 also reports the per-

**Table 7** Salient object detection performance on the MSO dataset. For the baseline USOD, we report its performance using the PR parameter that gives the optimal F-measure (%). We also report the performance of each method on a subset of the MSO dataset, which only contain images with salient objects (see Obj. Img. below).

		Prec.	Rec.	F-score
Full Dataset	USOD	77.5	74.0	75.7
	USOD+SOS	79.6	79.5	79.5
	USOD+GT	83.9	81.7	82.8
Obj. Img.	USOD	78.0	81.0	79.4
	USOD+SOS	79.5	81.8	80.6
	USOD+GT	83.9	81.7	82.8

mance of each method on images with salient objects. On this subset of images, using SOS improves the baseline USOD by about 1 absolute percentage point in F-measure. This suggests that our CNN SOS model is not only helpful for suppressing detections on background images, but is also beneficial by determining the number of detection windows for images with salient object.

**Cross-dataset generalization test for identifying background images.** Detecting background images is also useful for tasks like salient region detection and image thumbnailing (Wang et al., 2012). To test how well the performance of our SOS model 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 by Wang et al. (2012). The test set used by Wang et al. (2012) is composed of 5000 thumbnail images from the Web, and 3000 images sampled from the MSRA-B Liu et al. (2011) dataset. 50% of these images contain a single salient object, and the rest contain no salient object. Images for MSRA-B are resized to  $130 \times 130$  to simulate thumbnail images (Wang et al., 2012).

In Table 8, we report the detection accuracy of our CNN SOS model, in comparison with the 5-fold cross-validation accuracy of the best model reported by Wang et al. (2012). Note that our SOS model is trained on a different dataset, while the compared model is trained on a subset of the tested dataset via cross validation. Our method outperforms the model of Wang et al. (2012), and it can give fast prediction without resorting to any salient object detection methods. In contrast, the model of Wang et al. (2012) requires computing several saliency maps, which takes over 4 seconds per image as reported by Wang et al. (2012).

**Table 8** Recognition accuracy in predicting the presence of salient objects on the thumbnail image dataset (Wang et al., 2012). We show the 5-fold cross validation accuracy reported in (Wang et al., 2012). While our method is trained on the MSO dataset, it generalizes well to this other dataset.

	Wang et al. (2012)	Ours
accuracy (%)	82.8	84.2

## 6.2 Image Retrieval

In this section, we show an application of SOS in Content Based Image Retrieval (CBIR). In CBIR, many search queries refer to object categories. It is useful in many scenarios that users can specify the number of object instances in the retrieved images. For example, a designer may search for stock images that contain two animals to illustrate an article about couple relationships, and a parent may want to search his/her photo library for photos of his/her baby by itself.

We design an experiment to demonstrate how our SOS model can be used to facilitate the image retrieval for number-object (*e.g.* “three animals”) search queries. For this purpose, we implement a tag prediction system. Given an image, the system will output a set of tags with confidence scores. Once all images in a database are indexed using the predicted tags and scores, retrieval can be carried out by sorting the images according to the confidence scores of the query tags.

**The tag prediction system.** Our tag prediction system uses 6M training images from the Adobe Stock Image website. Each training image has 30-50 user provided tags. We pick about 18K most frequent tags for our dictionary. In practice, we only keep the first 5 tags for an image as we empirically find that first few tags are usually more relevant. Noun Tags and their plurals are merged (*e.g.* “person” and “people” are treated as the same tag). We use a simple KNN-base voting scheme to predict image tags. Given a test image and a Euclidean feature space, we retrieve the 75 nearest neighbors in our training set using the distance encoded product quantization scheme of Heo et al. (2014). The proportion of the nearest neighbors that have a specific tag is output as the tag’s confidence score. The Euclidean feature space for the KNN system is learned by a CNN model. We use the GoogleNet architecture and use the last 1024D average pooling layer as our feature space. Details about the CNN feature embedding training are included in the supplementary material.

**Dataset.** We use the public available NUS-WIDE dataset as our test set (Chua et al., 2009), which contains about 270K images. We index all the images of NUS-WIDE using our tag prediction system for all the

tags of our dictionary. The NUS-WIDE dataset has the annotation of 81 concepts, among which we pick all the concepts that correspond to countable object categories as our base test queries (see Fig. 12 for the 37 chosen concepts). For a base test query, say “animal”, we apply different test methods to retrieve images for four sub-queries, “one animal”, “two animals”, “three animals” and “many animals”, respectively. Then all the retrieved images for “animal” by different test methods are mixed together for annotation. We ask three subjects to label each retrieved image as one of the four sub-queries or none of the sub-queries (namely a five-way classification task). The subjects have no idea which test method retrieved which image. Finally, the annotations are consolidated by majority vote to produce the ground truth for evaluation.

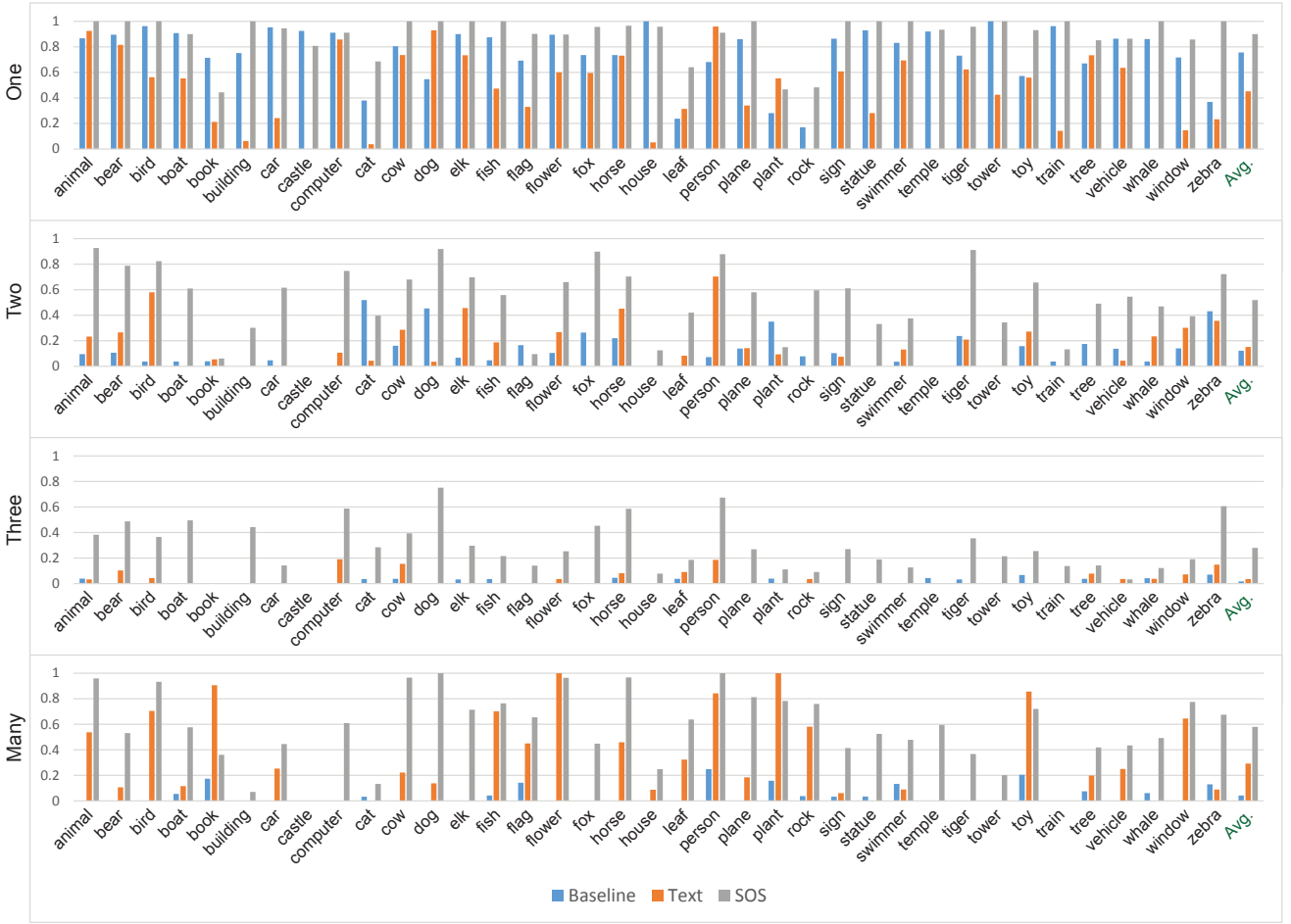
**Methods.** Given the tag confidence scores of each image by our tag prediction system, we use different methods to retrieve images for the number-object queries.

- Baseline. The baseline method ignores the number part of a query, and retrieves images using only the object tag.
- Text-based method. This method treats each sub-query as the combination of two normal tags. Note that both the object tags and the number tags are included in our dictionary. We multiply the confidence scores of the object tag with the confidence scores of the number tag (“one”, “two”, “three” or “many”). Then the top images are retrieved according to the multiplied scores.
- SOS-based method. This method differs from the text-based method in that it replaces the number tag confidence score with the corresponding SOS confidence score. For a number tag “one/two/three/many”, we use the SOS confidence score for 1/2/3/4+ salient object(s).

**Evaluation Metric.** The widely used Average Precision (AP) requires annotation of the whole dataset for each number-object pair, which is too expensive. Therefore, we use the normalized Discounted Cumulative Gain (nDCG) metric, which only looks at the top retrieved results. The nDCG is used in a recent image retrieval survey paper by Li et al. (2016) for benchmarking various image retrieval methods. The nDCG is formulated as

$$nDCG_h(t) = \frac{DCG_h(t)}{IDCG_h(t)}, \quad (2)$$

where  $t$  is the test query,  $DCG_h(t) = \sum_{i=1}^h \frac{2^{rel_i} - 1}{\log_2(i+1)}$ , and  $rel_i$  denotes the tag relevance of the retrieved image at position  $i$ . In our case,  $rel_i$  is either 0 or 1. The



**Fig. 12** nDCG scores for compared methods. For each object class, we use different methods to retrieve images of one/two/three/many object(s) of such class. The last column shows the average nDCG scores across different object classes.

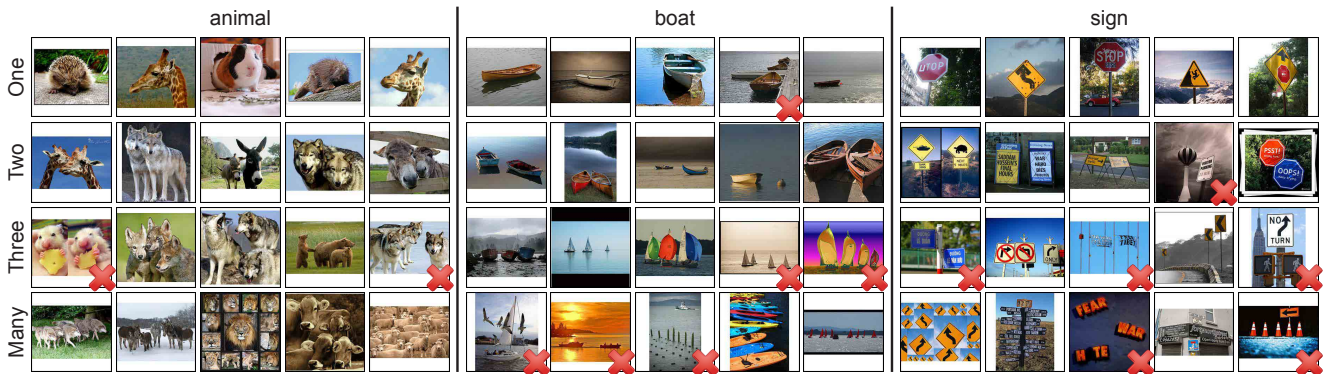
$IDCG_h(t)$  is the maximum possible  $DCG$  up to position  $h$ . We retrieve 20 images for each method, so we set  $h = 20$  and assume that there are at least 20 relevant images for each query.

**Results.** The nDCG scores of our SOS-based method, the text-based method and the baseline method are reported in Fig. 12. The SOS-based method gives consistently better average nDCG scores across queries for different numbers of objects, especially for the queries for more than one object. The scores of the SOS-based method for the group “three” are overall much lower than for the other groups. This is because the accuracy of our SOS is relatively lower for three objects. Moreover, there are many object categories that lack images with three objects, *e.g.* “statue”, “rock”, *etc.*

The baseline method gives pretty good nDCG scores for a single object, but for the other number groups, its performance is the worst. This reflects that images retrieved by a single object tag tend to contain only one dominant object. Note that it is often favorable that

the retrieved images present a single dominant object of the searched category when no number is specified. When using SOS, the performance in retrieving images of one object is further improved, indicating it can be beneficial to apply SOS by default for object queries.

The text-based method is significantly worse than our SOS-based method across all number groups. We observe that when a query has a number tag like “one”, “two” and “three”, the retrieved images by the text-based method tends to contain the given number of people. We believe that this is because these number tags often refer to the number of people in our training images. This kind of data bias obstructs the simple text-based approach to handling number-object queries. In contrast, our SOS-based method can successfully retrieve images for a variety of number-object queries thanks to the category agnostic nature of our SOS formulation. Sample results of our SOS-based method are shown in Fig. 13.



**Fig. 13** Sample results of the SOS-based method for number-object image retrieval. The base object tags are shown above each block. Each row shows the top five images for a number group (one/two/three/many). Irrelevant images are marked by a red cross.

## 7 Conclusion

In this work, we formulate the Salient Object Subitizing (SOS) 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 collect an SOS image dataset, and present a Convolutional Neural Network (CNN) model for this task. We leverage simple synthetic images to improve the CNN model training. Extensive experiments are conducted to show the effectiveness and generalizability of our CNN-based SOS method. We visualize that the features learned by our CNN model capture generic visual patterns that are useful for subitizing, and show how our model can generalize to unseen object categories. The usefulness of SOS is demonstrated in unconstrained salient object detection and content-based image retrieval. We show that our SOS model can improve the state-of-the-art salient object detection method, and it provides an effective solution to retrieving images by number-object queries.

## Acknowledgments

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

## References

- Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- Dwi Anoraganingrum. Cell segmentation with median filter and mathematical morphology operation. In *International Conference on Image Analysis and Processing*, 1999.

- Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Interactive object counting. In *European Conference on Computer Vision (ECCV)*, 2014.
- Janette Atkinson, Fergus W Campbell, and Marcus R Francis. The magic number  $4 \pm 0$ : A new look at visual numerosity judgements. *Perception*, 5(3):327–34, 1976.
- Tamara L Berg and Alexander C Berg. Finding iconic images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2009.
- Ali Borji, Dicky N Sihite, and Laurent Itti. Salient object detection: A benchmark. In *European Conference on Computer Vision (ECCV)*, 2012.
- Sarah T Boysen and E John Capaldi. *The development of numerical competence: Animal and human models*. Psychology Press, 2014.
- Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- Antoni B Chan, Z-SJ Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- 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.
- Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.
- Jiwon Choi, Chanhoo Jung, Jaeho Lee, and Changick Kim. Determining the existence of objects in an image and its application to image thumbnailing. *Signal Processing Letters*, 21(8):957–961, 2014.
- Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.
- Douglas H Clements. Subitizing: What is it? why teach it? *Teaching children mathematics*, 5:400–405, 1999.
- Hank Davis and Rachelle Pérusse. Numerical competence in animals: Definitional issues, current evidence, and a new



- research agenda. *Behavioral and Brain Sciences*, 11(04): 561–579, 1988.
- Stanislas Dehaene. *The number sense: How the mind creates mathematics*. Oxford University Press, 2011.
- 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>, 2007.
- Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
- Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun. Salient object detection by composition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- Viswanath Gopalakrishnan, Yiqun Hu, and Deepu Rajan. Random walks on graphs to model saliency in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- Hans J Gross. The magical number four: A biological, historical and mythological enigma. *Communicative & Integrative Biology*, 5(1):1–2, 2012.
- Hans J Gross, Mario Pahl, Aung Si, Hong Zhu, Jürgen Tautz, and Shaowu Zhang. Number-based visual generalisation in the honeybee. *PloS one*, 4(1):e4263, 2009.
- Jae-Pil Heo, Zhe Lin, and Sung-Eui Yoon. Distance encoded product quantization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *Workshop on Deep Learning, NIPS*, 2014.
- Brenda RJ Jansen, Abe D Hofman, Marthe Straatemeier, Bianca MCW Bers, Maartje EJ Raijmakers, and Han LJ Maas. The role of pattern recognition in children’s exact enumeration of small numbers. *British Journal of Developmental Psychology*, 32(2):178–194, 2014.
- W Stanley Jevons. The power of numerical discrimination. *Nature*, 3:281–282, 1871.
- Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014.
- EL Kaufman, MW Lord, TW Reese, and J Volkmann. The discrimination of visual number. *The American Journal of Psychology*, pages 498–525, 1949.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.
- Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.
- Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys*, 49(1):14:1–14:39, June 2016.
- Yin Li, Xiaodi Hou, Christof Koch, J Rehg, and A Yuille. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.
- Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.
- George Mandler and Billie J Shebo. Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, 111(1):1, 1982.
- Sumit K Nath, Kannappan Palaniappan, and Filiz Buncyak. Cell segmentation using coupled level sets and graph-vertex coloring. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2006.
- Mario Pahl, Aung Si, and Shaowu Zhang. Numerical cognition in bees and other insects. *Frontiers in psychology*, 4, 2013.
- Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.
- Manuela Piazza and Stanislas Dehaene. From number neurons to mental arithmetic: The cognitive neuroscience of number sense. *The Cognitive Neurosciences*, 3rd edition, pages 865–77, 2004.
- Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Deep Vision Workshop*, 2014.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- Christian Scharfenberger, Steven L Waslander, John S Zelek, and David 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.
- Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2014.
- Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,

2013.

Michael Stark, Michael Goesele, and Bernt Schiele. Back to the future: Learning shape models from 3D CAD data. In *British Machine Vision Conference (BMVC)*, 2010.

Ivilin Stoianov and Marco Zorzi. Emergence of a visual number sense in hierarchical generative models. *Nature neuroscience*, 15(2):194–196, 2012.

Venkatesh Bala Subburaman, Adrien Descamps, and Cyril Carincotte. Counting people in the crowd using a generic head detector. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2012.

Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *British Machine Vision Conference (BMVC)*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Antonio Torralba, Kevin P Murphy, William T Freeman, and Mark A Rubin. Context-based vision system for place and object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

Lana M Trick and Zenon W Pylyshyn. Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological review*, 101(1): 80, 1994.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/>, 2008.

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

Peng Wang, Jingdong Wang, Gang Zeng, Jie Feng, Hongbin Zha, and Shipeng Li. Salient object detection for searched web images via global saliency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

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

Jianming Zhang, Shuga Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. Salient object subitizing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015a.

Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. Minimum barrier salient object detection at 80 fps. In *IEEE International Conference on Computer Vision (ICCV)*, 2015b.

Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. Unconstrained salient object detection via proposal subset optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Will Y. Zou and James L. McClelland. Progressive development of the number sense in a deep neural network. In *Annual Conference of the Cognitive Science Society (CogSci)*,

2013.