Robust Multi-Object Detection Based on Data Augmentation with Realistic Image Synthesis for Point-of-Sale Automation

Saiprasad Koturwar, Soma Shiraishi, Kota Iwamoto
Data Science Research Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211-8666, Japan

Abstract

As an alternative to bar-code scanning, we are developing a real-time retail product detector for point-of-sale automation. The major challenge associated with image based object detection arise from occlusion and the presence of other objects in close proximity. For robust product detection under such conditions, it is crucial to train the detector on a rich set of images with varying degrees of occlusion and proximity between the products, which fairly represents a wide range of customer tendencies of placing products together. However, generating a fairly large database of such images traditionally requires a large amount of human effort. On the other hand, acquiring individual object images with their corresponding masks is a relatively easy task. We propose an realistic image synthesis approach which uses individual object images and their corresponding masks to create training images with desired properties (occlusion and congestion among the products). We train our product detector over images thus generated and achieve a consistent performance improvement across different types of test data. With the proposed approach, detector achieves an improvement of 46.2% (from 0.67 to 0.98) and 40% (from 0.60 to 0.84) over precision and recall respectively, compared to using a basic training dataset containing one product per image.

1 Introduction

Over past few years, retail industry has been seeking solutions for facilitating unmanned store operations, owing to labor shortages and longer queue times. In particular, executing an automated checkout process is crucial, since it has been the bottleneck for smooth operation of stores. Here we instantiate the automated checkout system by replacing conventional bar-code scanning with an image recognition based product detector system. Image recognition enables scanning of multiple products at once, making the checkout faster. This system leverages state-of-the-art deep learning to enable reliable recognition of various types of products including packaged products (e.g. pre-processed foods, soft-drinks etc.). The developed POS system is shown in Fig. 1. The process for checkout is as follows. First, the customer puts the products he/she wishes to buy on the product placement platform. Next, a camera placed above captures an image of the products placed on the platform. Finally, the system analyzes the captured image to recognize individual products, which then outputs the total price on the screen.

Figure 1: The developed POS system.

Figure 2: Examples of difficult instances for detection.

To enable efficient and user friendly operation, with minimal constraints on the user, we allow the users to place products freely and randomly as long as they do not occlude each other heavily. This freedom of product placement results in congested conditions as shown in Fig. 2. In most cases, the products are placed in close proximity where sometimes they are touching each other or even have partial occlusion. Hence, it is vital to have a robust identification of products under such congested conditions in order to have flawless operation of the proposed system.

In order to achieve a reliable product recognition under such congested conditions, our system divides the recognition task into two stages: *detection* and *classification*. The detector focuses on robustly localizing and extracting the regions of each individual product from the image. The detector then passes each extracted product region to the classifier, which focuses on the classification of the product (out of hundreds of possible product classes) one-by-one. This

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

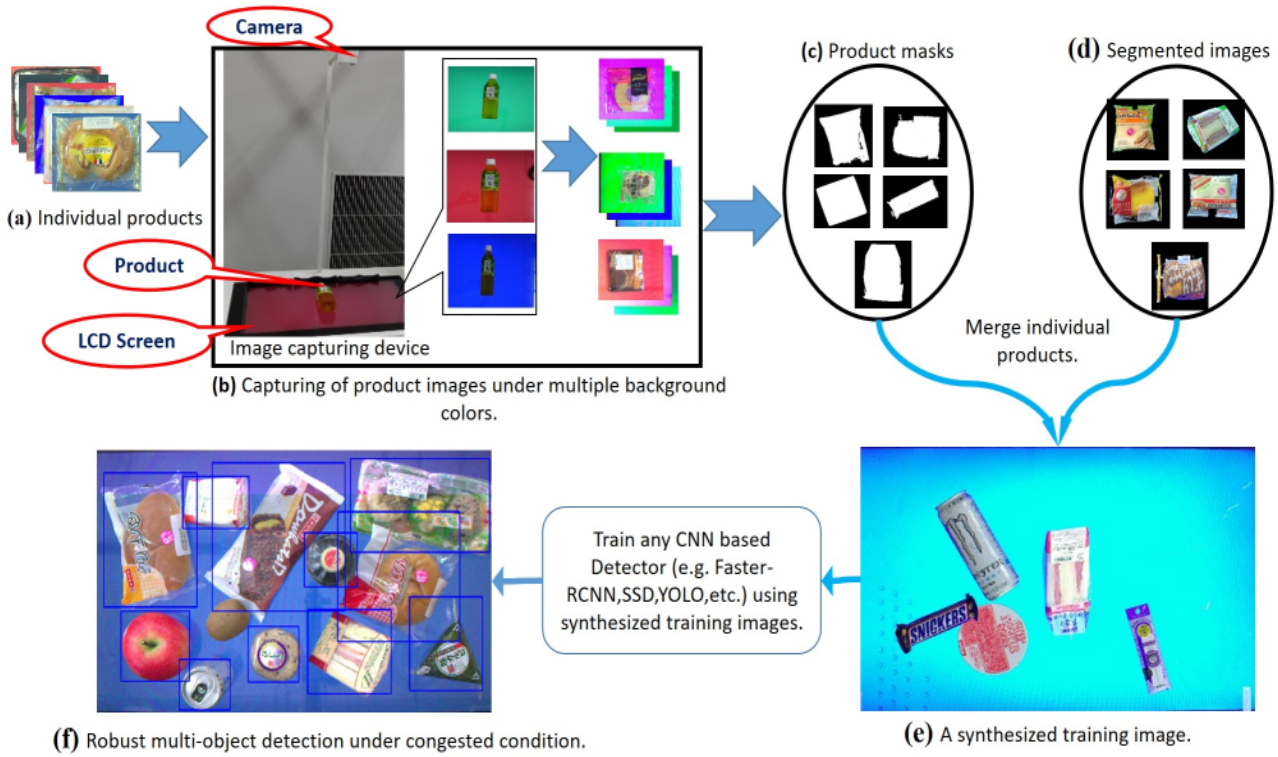


Figure 3: Overall process flow of the proposed realistic image synthesis used for training the product detector.

way, the classifier does not need to be robust to occlusion and other products in close proximity, as opposed to state-of-the-art recognition approaches such as Faster-RCNN(Ren et al. 2015), SSD(Liu et al. 2016) and YOLO (Redmon et al. 2015), which perform an integrated detection and classification step.

In the above setting, once the detector robustly detects individual products under congested conditions, the classification task becomes easier. When we train state-of-the-art object detectors on sample images of individual products, we fail to achieve good performance on instances shown in Fig. 2. In order to have robust retail product detection in real life, it is crucial that we train our detector on images that are representative of product placement resulting from normal human behavior, such as products placed in a congested manner. One way to gather such training data is by manually placing the products and capturing those images. However, this approach is laborious and time consuming.

In this work, we propose an approach for data augmentation by efficiently synthesizing realistic training images using individual product images along with their corresponding masks (Fig. 3). As a result of this approach, we are able to reduce the human labor required for data acquisition. Although an independent work done in (Follmann, Drost, and Böttger 2018) proposes a similar approach for generating training images using individual product masks, they use a standard background color subtraction approach for generating product masks. We found that, this approach is not robust and is sensitive to background color. In contrast, we propose a color-insensitive, robust mask generation system, which we describe in section 3.1. Furthermore, they do not have control over the relative placement of products in resulting images, which is central to our proposed method and is important for our intended use. In our approach, we control proximity and occlusion in resultant images through a parameter which we name as overlap index. Experiments reveal that this particular parameter is crucial and needs to be tuned. We tune the overlap index to generate realistic training images and achieve an improvement of 46.2% on precision and 40% on recall compared to using a basic training dataset containing one product per image. We would like to note that this work only focuses on detection (localization) of products and not on the classification of products, which we treat as a separate problem.

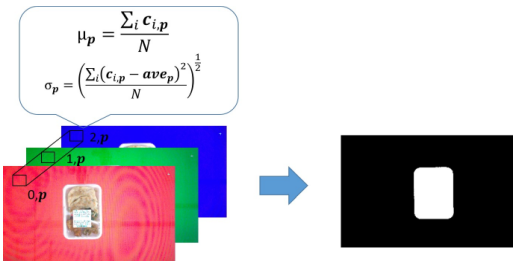


Figure 4: Mask extraction by thresholding standard deviation image.

Figure 5: Merging individual products. *Left* :As the newly added product overlaps with the existing product/s, move the new product away from the existing product/s, *right*: As the newly added product does not overlap with the existing product/s, move the new product towards the existing product/s.

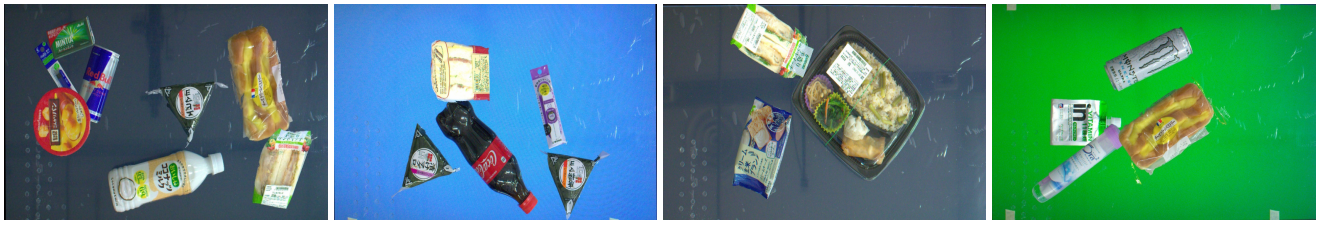


Figure 6: Sample training images generated by the proposed realistic image synthesis with overlap index=0.01.

2 Related Work

In this section, we review the related works that tackle object detection for general purpose. Historically, focus has been on extracting handcrafted features such as SIFT (Lowe 2004), HoG (Dalal and Triggs 2005) for object detection. These approaches, although useful in many contexts, are not successful in detecting products with amorphous textures (such as fruits or non-packaged products). On the other hand, CNNs are capable of modeling various complex patterns present in the images. Therefore, recently the focus has shifted towards developing sophisticated CNN architectures such as Faster-RCNN, YOLO and SSD. However, as discussed in the previous section, these CNN based approaches, do not explicitly tackle the problem of occlusion and presence of objects in proximity, but concentrate on developing a sophisticated architecture for object detection.

Another work discussed in (Qiao et al. 2017) concentrates on object proposal for retail product detection in supermarkets. They tackle the problem of proximity and occlusion by integrating the object scale prediction in detection framework. As with the previous object detectors, this approach also proposes a sophisticated architecture for object detection and requires large training data for good performance. Unlike these approaches which focus on improvement of the CNN structure, we tackle the detection under occlusion and proximity using data augmentation. In particular, we synthesize training images that are representative of the task at hand, which in this case is the product placement in congested manner.

a) Non-augmented b) Simple cut and paste

Figure 7: Sample training images of the approaches used for ablation study.

3 Realistic Image Synthesis

As discussed before, it is time consuming and laborious to acquire large amounts of training images having multiple products placed in a congested manner, which is the result of natural usage of our POS system. We propose an image synthesis approach to automatically create these training images by merging product images taken individually, which dramatically reduces the human efforts required for gathering training images. Viability of the proposed approach depends on how realistic (indistinguishable from actual captured images) the resulting images are. In order to achieve realism in images, our proposal consists of two crucial stages: *precise mask extraction* of individual products and *merging individual products*. Figure 3 explains the steps involved in realistic image synthesis. Following sections explain these steps in detail.

3.1 Precise Mask Extraction

In order to have realism, it is important that the generated images do not have any unrealistic edges or artifacts. Thus, it is vital to have the exact mask information of individual products. However, automatically extracting an accurate mask is not a trivial task. Traditional approaches, such as using a fix-colored background (in most cases, green) does not yield accurate mask extraction. Especially, the approach fails when the product whose mask we are trying to extract shares its color with the background. We propose a mask extraction method which is more robust than the traditional approach and is also very easy to automate.

Table 1: Training dataset.(OI is overlap index)

Image Synthesis Approach	No. of images
Non-augmented	1,638
Simple cut and paste	20,000
Proposed (OI = 0.0, no overlap)	20,000
Proposed (OI = 0.01, optimum overlap)	20,000
Proposed (OI = 0.1, high overlap)	20,000

We make the mask generation approach robust to background color by capturing the product image under multiple background colors. To achieve this, we have built an image capturing device with LCD screen as the placement platform (Fig. 3 (b)), whose color can be changed. In order to generate precise masks, we place individual products on the LCD screen and capture it under multiple background colors. Once we have a set of images for a single product with multiple background colors (Fig. 4), we take the pixel-wise standard deviation across all background colors, and generate a standard deviation image (M_{std}),

$$\mu_p = \frac{\sum_i c_{i,p}}{N}$$

$$\sigma_p = \left(\frac{\sum_i (c_{i,p} - u_p)^2}{N}\right)^{\frac{1}{2}}$$

where, N is the number of images captured with different background, $c_{i,p}$ is the image intensity at pixel p in image i, μ_p and σ_p are the average and the standard deviation of pixel values at pixel p across images of different color backgrounds for a given product, respectively.

Foreground pixels in the generated standard deviation image M_{std} should have a lower value, as changing background color does not affect the pixel values of foreground. Thus, binary image resulting after the thresholding operation on M_{std} is a precise mask of the product. A precise product mask is necessary for generating realistic training images through merging, which is explained in the next section.

3.2 Merging Individual Products

A naive way of merging individual products would be by randomly pasting the chosen products on a given background. However, in this approach, we do not have control over the arrangement of products in the resulting images. We want to synthesize training images which are representative of the product placement resulting from natural usage of our POS system, i.e. having multiple products placed in a congested manner. We achieve controllability over the extent of occlusion and proximity in the resulting images by controlling *overlap index*. Overlap index is the maximum amount of overlap allowed between two individual products, which is calculated as the IoU (intersection over union) between the product masks. A larger overlap index means the resulting images will have more overlap between the products. The exact process for synthesizing these images using the overlap index is summarized below.

We first fix the overlap index before synthesizing training images. Next, we randomly sample multiple products from the pool of available training products. Out of the sampled products, we choose one product at random and paste it on an empty image, which we refer to as the base image. Finally, we add each subsequent product in the base image using its mask information. We move the newly added product based on it's overlap index such that the maximum overlap is equal to the preset value (Fig. 5).

Overlap index has significant impact on resulting images. As we increase the overlap index, resulting images contain many occluded instances. This aggressive occlusion may not be ideal for our particular case. On the other hand, when we restrict overlap index to zero, it results in images with no overlap. However, images synthesized with a small overlap index (Fig. 6) are representative of the images encountered in the natural usage of our POS system, thus will be an optimum choice for training the object detector.

4 Experiments and Results

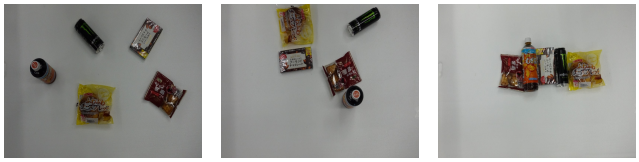
4.1 Ablation Study Approaches

In order to show the effectiveness of our proposed realistic image synthesis approach for generating training images, we perform ablation experiment by comparing our method against other training image synthesis approaches, specifically, non-augmented approach and simple cut and paste approach. In the non-augmented approach, training dataset is generated by manually capturing images of individual products under different orientations. The training images obtained using this approach have realism but lack congestion, as shown in Fig. 7(a). In the simple cut and paste approach, the products are cropped along the bounding box from corresponding individual image, and then merged by keeping fair proximal distance. This approach results in congested cases, but it lacks realism as the unrealistic edges are introduced, as shown in Fig. 7(b).

4.2 Experimental Setup

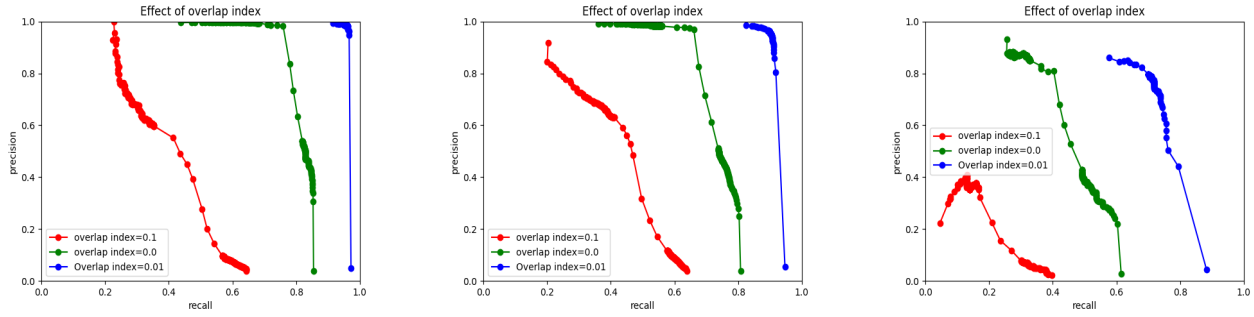
Training dataset: All the training image synthesis approaches discussed in this work need individual product images as the seed. We use a total of 293 products as the seed for synthesizing training images. The total number of synthesized images for each approach is summarized in Table 1. For the non-augmented approach, we take a single product and capture images under different orientations, typically ranging from 5 to 7 orientations per product.

Test dataset: Test dataset contains real (not synthesized) images captured by the POS system shown in Fig. 1. Each image consists of multiple products which do not appear



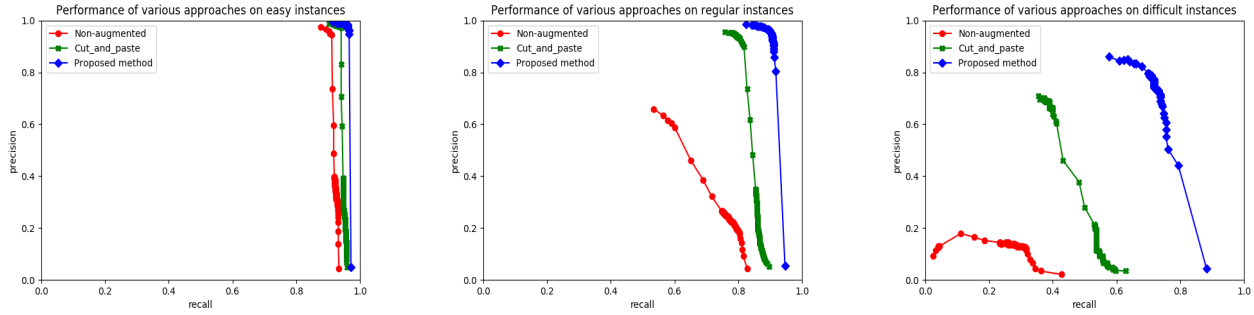
a) Easy dataset b) Regular dataset c) Difficult dataset

Figure 8: Test datasets with varying level of difficulty.



a) Easy dataset b) Regular dataset c) Difficult dataset

Figure 9: PR curves for different values of overlap index on the proposed realistic image synthesis approach.



a) Easy dataset b) Regular dataset c) Difficult dataset

Figure 10: PR curves for different training image synthesis approaches.

in the training dataset. We categorize the dataset into three groups with varying level difficulty based on proximity among products: easy (total of 300 images), regular (total of 600 images) and difficult (total of 100 images), as shown in Fig. 8. Easy dataset contains images where individual products are fairly separated from each other. Regular dataset contains images where individual products are placed at close proximity, with a little overlap between adjacent products. This dataset is representative of images we expect to encounter often while using our POS system. Difficult dataset contains images where individual products are placed at lowest possible proximity. Total number of individual products used for the test dataset is 153.

4.3 Results and Performance Analysis

We evaluate the object detection performance of individual approaches on the three test datasets. We use precision, recall, and average overlap over ground truth boxes (AOGB) as evaluation criterion. The prediction is regarded as correct when the IoU between the predicted bounding box and the groundtruth bounding box is greater than 0.5. The precision and the recall are defined as, $precision = \frac{N}{P}$, $recall = \frac{N}{T}$, where N is the number of correct predictions, P is the total number of predictions, and T is the total number of groundtruth instances. We define AOGB as the average IoU of predicted bounding box with ground truth bounding box across correct predictions. We chose Faster-RCNN as our object detector as it achieves best performance on MS-COCO dataset, and ResNet-101 as feature extractor with learning rate of 0.003. We individually train the detector end-to-end for 200K steps for all approaches in Table 1.

To see the effect of the overlap index in the proposed method on the performance of the object detector, we experiment with different values of overlap index. Figure 9 shows the PR (precision vs recall) curves obtained by changing the confidence threshold for predictions. Table 2 shows the figures for a fixed confidence threshold of 0.8. The results show that the selection of the overlap index has a significant impact on the performance. We achieve the optimal precision and recall of (0.98,0.84) on the regular dataset for the value of overlap index = 0.01. Using the optimum overlap index of 0.01, we compare the proposed approach against other approaches. Figure 10 shows the PR curves comparing the different training image synthesis approaches. Table 3 shows the figures for a fixed confidence threshold of 0.8. Proposed method outperforms other approaches in terms of precision and recall across all levels of difficulty. Specifically, on the regular dataset, the proposed method achieves a precision and recall of (0.98,0.84), which is a significant improvement over (0.92,0.80) for cut and paste approach and (0.67,0.60) for non-augmented approach. Performance gain is more evident for the difficult dataset, where our proposed method

Table 2: Performance across different values of overlap index at confidence threshold of 0.8.

	Easy dataset			Regular dataset			Difficult dataset		
Overlap index	Recall	Precision	AOGB	Recall	Precision	AOGB	Recall	Precision	AOGB
0.0	0.71	0.98	0.83	0.61	0.98	0.82	0.36	0.83	0.75
0.01	**0.93**	**0.99**	**0.86**	**0.84**	**0.98**	**0.85**	**0.61**	**0.84**	**0.78**
0.1	0.41	0.55	0.73	0.44	0.59	0.69	0.16	0.36	0.68

Table 3: Performance across different training image synthesis approaches at confidence threshold of 0.8.

	Easy dataset			Regular dataset			Difficult dataset		
Approach	Recall	Precision	AOGB	Recall	Precision	AOGB	Recall	Precision	AOGB
Non - Augmented	0.89	0.97	0.84	0.60	0.67	0.76	0.11	0.03	0.60
Cut and paste	0.94	0.97	0.84	0.80	0.92	0.81	0.4	0.64	0.736
Proposed	**0.93**	**0.99**	**0.86**	**0.84**	**0.98**	**0.85**	**0.61**	**0.84**	**0.78**

achieves a precision and recall of (0.84,0.61) as compared to (0.64,0.4) for cut and paste approach and (0.03,0.11) for non-augmented approach.

It is important to note that for a high overlap index of 0.1, the overall performance of the detector is worse than both simple cut and paste approach and non-augmented approach. This is because an aggressive merging results in excessive occluded training instances, unlikely to be encountered using our POS system. On the other hand, if we constrain the images to not have any occlusion, we achieve a performance gain over the cut-and-paste approach (Table 2), but is not robust enough to occlusion and congestion. In our case, we expect the test data to have congestion, but not full occlusion. Thus, when we restrict the overlap index to 0.01, we achieve the best performance across all approaches, as these training images have optimal congestion as well as mild occlusion among the products.

5 Conclusion

In this work, we proposed a data augmentation method by synthesizing realistic training images to enable robust retail product detection under congested conditions for POS automation. The proposed method consists of two crucial steps, *precise mask extraction* and *merging of individual products*. The proposed mask extraction method can be automated and is more accurate than conventional approaches. Furthermore, the proposed product merging approach gives us the control over the distribution of synthesized training images through overlap index. We have studied the effect of overlap index i.e. how close the products are placed. Our study reveals that by choosing the optimal value for overlap index, the object detector trained using images synthesized by the proposed approach achieves a robust detection under the congested conditions. The experimental results show that the proposed approach achieves significant improvement of 46.2% (from 0.67 to 0.98) and 40% (from 0.60 to 0.84) over precision and recall respectively, compared to using a basic non-augmented training dataset containing one product per image.

References

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, 886–893. IEEE Computer Society.

Follmann, P.; Drost, B.; and Böttger, T. 2018. Acquire, augment, segment enjoy: Weakly supervised instance segmentation of supermarket products. *ArXiv e-prints*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *ECCV*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60(2):91–110.

Qiao, S.; Shen, W.; Qiu, W.; Liu, C.; and Yuille, A. 2017. ScaleNet: Guiding Object Proposal Generation in Supermarkets and Beyond. *ICCV 2017*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2015. You Only Look Once: Unified, Real-Time Object Detection. *CVPR 2016*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS'15, 91–99. MIT Press.