Precise Detection in Densely Packed Scenes

Eran Goldman\textsuperscript{1,3}\footnote{Equal Contribution.}  Roei Herzig\textsuperscript{2}  Aviv Eisenschtat\textsuperscript{3}\footnote{Equal Contribution.}  Jacob Goldberger\textsuperscript{1}  Tal Hassner\textsuperscript{4}\footnote{Work done while at the University of Southern California.}

\textsuperscript{1}Bar-Ilan University, \textsuperscript{2}Tel Aviv University, \textsuperscript{3}Trax Retail, \textsuperscript{4}The Open University of Israel

Abstract

Man-made scenes can be densely packed, containing numerous objects, often identical, positioned in close proximity. We show that precise object detection in such scenes remains a challenging frontier even for state-of-the-art object detectors. We propose a novel, deep-learning based method for precise object detection, designed for such challenging settings. Our contributions include: (1) A layer for estimating the Jaccard index as a detection quality score; (2) a novel EM merging unit, which uses our quality scores to resolve detection overlap ambiguities; finally, (3) an extensive, annotated data set, SKU-110K, representing packed retail environments, released for training and testing under such extreme settings. Detection tests on SKU-110K and counting tests on the CARPK and PUCPR+ show our method to outperform existing state-of-the-art with substantial margins. The code and data will be made available on www.github.com/eg4000/SKU110K_CVPR19.

1. Introduction

Recent deep learning–based detectors can quickly and reliably detect objects in many real world scenes \cite{girshick2014rich, redmon2016you, ren2015faster, liu2016ssd, chen2017rethinking, li2017ssd, zhang2018sharable}. Despite this remarkable progress, the common use case of detection in crowded images remains challenging even for leading object detectors.

We focus on detection in such densely packed scenes, where images contain many objects, often looking similar or even identical, positioned in close proximity. These scenes are typically man-made, with examples including retail shelf displays, traffic, and urban landscape images. Despite the abundance of such environments, they are under-represented in existing object detection benchmarks. It is therefore unsurprising that state-of-the-art object detectors are challenged by such images.

To understand what makes these detection tasks difficult, consider two identical objects placed in immediate proximity, as is often the case for items on store shelves (Fig. 1).

The challenge is to determine where one object ends and the other begins; minimizing overlaps between their adjacent bounding boxes. In fact, as we show in Fig. 1(a,c), the state-of-the-art RetinaNet \cite{lazebnik2018deformable}, often returns bounding boxes which partially overlap multiple objects or detections of adjacent object regions as separate objects.

We describe a method designed to accurately detect objects, even in such densely packed scenes (Fig. 1(b,d)). Our method includes several innovations. We propose learning the Jaccard index with a soft Intersection over Union (Soft-IoU) network layer. This measure provides valuable information on the quality of detection boxes. We explain how detections can be represented as a Mixture of Gaussians (MoG), reflecting their locations and their Soft-IoU.
scores. An Expectation-Maximization (EM) based method is then used to cluster these Gaussians into groups, resolving detection overlap conflicts.

To summarize, our novel contributions are as follows:

- **Soft-IoU layer**, added to an object detector to estimate the Jaccard index between the detected box and the (unknown) ground truth box (Sec. 3.2).

- **EM-Merger unit**, which converts detections and Soft-IoU scores into a MoG, and resolves overlapping detections in packed scenes (Sec. 3.3).

- **A new data set and benchmark**, the store keeping unit, 110k categories (SKU-110K), for item detection in store shelf images from around the world (Sec. 4).

We test our detector on SKU-110K. Detection results show our method to outperform state-of-the-art detectors. We further test our method on the related but different task of object counting, on SKU-110K and the recent CARPK and PUCPR+ car counting benchmarks [22]. Remarkably, although our method was not designed for counting, it offers a considerable improvement over state-of-the-art methods.

### 2. Related work

**Object detection.** Work on this problem is extensive and we refer to a recent survey for a comprehensive overview [29]. Briefly, early detectors employed sliding window–based approaches, applying classifiers to window contents at each spatial location [10, 46]. Later methods narrow this search space by determining region proposals before applying sophisticated classifiers [1, 7, 31, 44, 51].

Deep learning–based methods now dominate detection results. To speed detection, proposal-based detectors such as R-CNN [15] and Fast R-CNN [16] were developed, followed by Faster R-CNN [38] which introduced a region proposal network (RPN), then accelerated even more by R-FCN [9]. Mask-RCNN [19] later added segmentation output and better detection pooling [38]. We build on these methods, claiming no advantage in standard object detection tasks. Unlike us, however, these two-stage methods were not designed for crowded scenes where small objects appear in dense formations.

Recently, some offered proposal-free detectors, including YOLO [36], SSD [30], and YOLO9000 [37]. To handle scale variance, feature pyramid network (FPN) [26] added up-scaling layers. RetinaNet [27] utilized the same FPN model, introducing a Focal Loss to dynamically weigh hard and easy samples for better handling of class imbalances that naturally occur in detection datasets. We extend this approach, introducing a new detection overlap measure, allowing for precise detection of tightly packed objects.

These methods use hard-labeled log-likelihood detections to produce confidences for each candidate image region. We additionally predict a Soft-IoU confidence score which represents detection bounding box accuracy.

**Merging duplicate detections.** Standard non-maximum suppression (NMS) remains a de-facto object detection duplicate merging technique, from Viola & Jones [45] to recent deep detectors [27, 37, 38]. NMS is a hand-crafted algorithm, applied at test time as post-processing, to greedily select high scoring detections and remove their overlapping, low confidence neighbors.

Existing NMS alternatives include mean-shift [10, 46], agglomerative [5], and affinity propagation clustering [31], or heuristic variants [4, 40, 23]. GossipNet [21] proposed to perform duplicate-removal using a learnable layer in the detection network. Finally, others bin IoU values into five categories [43]. We instead take a probabilistic interpretation of IoU prediction and a very different general approach.

Few of these methods showed improvement over simple, greedy NMS, with some also being computationally demanding [21]. In densely packed scenes, resolving detec-
tion ambiguities is exacerbated due to the many overlapping detections. We propose an unsupervised method, designed for clustering duplicate detection in cluttered regions.

**Crowded scene benchmarks.** Many benchmarks were designed for testing object detection or counting methods and we survey a few in Table 1. Importantly, we are unaware of detection benchmarks intended for densely packed scenes, such as those of interest here.

Popular object detection sets include ILSVRC [12], PASCAL VOC [13] detection challenges, MS COCO [28], and the very recent Open Images v4 [25]. None of these provides scenes with packed items. A number of recent benchmarks emphasize crowded scenes, but are designed for counting, rather than detection [2, 8, 34].

As evident from Table 1, our new SKU-110K dataset, described in Sec. 4, provides one to three orders of magnitude more items per image than nearly all these benchmarks (the only exception is the PUCPR+ [22] which offers two orders of magnitude fewer images, and a single object class to our more than 110k classes). Most importantly, our enormous, per image, object numbers imply that all our images contain very crowded scenes, which raises the detection challenges described in Sec. 1. Moreover, identical or near identical items in SKU-110K are often positioned closely together, making detection overlaps a challenge. Finally, the large number of classes in SKU-110K implies appearance variations which add to the difficulty of this benchmark, even in challenges of object/non-object detection.

### 3. Deep IoU detection network

We build on a standard detection network design, described in Sec. 3.1. We extend this design in two ways. First, we define a novel Soft-IoU layer which estimates the overlap between predicted bounding boxes and the (unknown) ground truth (Sec. 3.2). These Soft-IoU scores are then processed by a proposed EM-Merger unit, described in Sec. 3.3, which resolves ambiguities between overlapping bounding boxes, returning a single detection per object.

#### 3.1. Base detection network

Our base detector is similar to existing methods [26, 27, 30, 38]. We first detect objects by building a FPN network [26] with three upscaling-layers, using ResNet-50 [20] as a backbone. The proposed model provides three fully-convolutional output heads for each RPN [38]: Two heads are standard and used also by previous work [27, 37] (our novel third head is described in Sec. 3.2).

The first is a *detection head* which produces a bounding box regression output for each object, represented as 4-tuples: \((x, y, h, w)\) for the 2D coordinates of a bounding box center, height and width. The second, *classification head* provides an *objectness score* (confidence) label, \(c \in [0, 1]\) (assuming an object/no-object detection task with one object class). In practice, we filter detections for which \(c \leq 0.1\), to avoid creating a bias towards noisy detections when training our Soft-IoU layer, described next.

#### 3.2. Soft-IoU layer

In non-dense scenes, greedy NMS applied to objectness scores, \(c\), can resolve overlapping detections. In dense images, however, multiple overlapping bounding boxes often reflect multiple, tightly packed objects, many of which receive high objectness scores. As we later show (Sec. 5.2), in such cases, NMS does not adequately discriminate between overlapping detections or suppress partial detections.

To handle these cluttered positive detections, we propose predicting an additional value for each bounding box: The IoU (i.e., Jaccard index) between a regressed detection box and the object location. This Soft-IoU score, \(c_{iou} \in [0, 1]\), is estimated by a fully-convolutional layer which we add as a third head to the end of each RPN in the detector.

Given \(N\) predicted detections, the IoU between a predicted bounding box \(b_i\), \(i \in \{1,..,N\}\) and its ground truth bounding box, \(\hat{b}_i\), is defined as:

\[
\text{IoU}_i = \frac{\text{Intersection}(\hat{b}_i, b_i)}{\text{Union}(\hat{b}_i, b_i)}.
\]
where \( n \) is therefore defined as:

Here, visualizing the output of the EM-Merger unit.

**Visualizing the output of the EM-Merger unit.** Raw detections on these images (not shown) contain many overlapping bounding boxes. Our approach of representing detections as a MoG (Eq. 5), visualized here as heat maps, provides clear signals for where items are located. The simplified MoG of Eq. (7) is visualized as green ellipsoids. See Sec. 3.3 for details.

Objects vs. Soft-IoU. The objectness score used in previous methods predicts object/no-object labels whereas our Soft-IoU predicts the IoU of a detected bounding box and its ground truth. So, for instance, a bounding box which partially overlaps an object can still have a high objectness score, \( c \), signifying high confidence that the object appears in the bounding box. For the same detection, we expect \( e^{\text{iou}} \) to be low, due to the partial overlap.

In fact, object/no-object classifiers are trained to be invariant to occlusions and translations. A good objectness classifier would therefore be invariant to the properties which our Soft-IoU layer is sensitive to. Objectness and Soft-IoU could thus be considered reflecting complementary properties of a detection bounding box.

### 3.3. EM-Merger unit for inference

We now have \( N \) predicted bounding box locations, each with its associated objectness, \( c \), and Soft-IoU, \( e^{\text{iou}} \), scores. Bounding boxes, especially in crowded scenes, often clump together in clusters, overlapping each other and their item locations. Our EM-Merger unit filters, merges, or splits these overlapping detection clusters, in order to resolve a single detection per object. We begin by formally defining these detection clusters.

**Detectors as Gaussians.** We consider the \( N \) bounding boxes produced by the network as a set of 2D Gaussians:

\[
F = \{f_i\}_{i=1}^N = \{\mathcal{N}(\mathbf{p}_i; \mu_i, \Sigma_i)\}_{i=1}^N, \tag{4}
\]

with \( \mathbf{p} \in \mathbb{R}^2 \), a 2D image coordinate. The \( i \)-th detection is thus represented by a 2D mean, the central point of the box, \( \mu_i = (x_i, y_i) \), and a diagonal covariance, \( \Sigma_i = [(h_i/4)^2, 0; 0, (w_i/4)^2] \), reflecting the box size, \( (h_i, w_i) \).

We represent these Gaussians, jointly, as a single Mixture of Gaussians (MoG) density:

\[
f(\mathbf{p}) = \sum_{i=1}^N \alpha_i f_i(\mathbf{p}), \tag{5}\]

where the mixture coefficients, \( \alpha_i = \frac{e^{\text{iou}_i}}{\sum_k e^{\text{iou}_k}} \), reflecting our confidence that the bounding box overlaps with its ground truth, are normalized to create a MoG.

Fig. 3 visualizes the density of Eq. (5) as heat-maps, translating detections into spatial region maps representing our per-pixel confidences of detection overlaps; each region weighted by the accumulated Soft-IoU.

**Selecting predictions: formal definition.** We next resolve our \( N \) Gaussians (detections) into precise, non-overlapping bounding box detections by using a MoG clustering method \([6, 17, 18, 50]\).

We treat the problem of resolving the final detections as finding a set of \( K < N \) Gaussians,

\[
G = \{g_j\}_{j=1}^K = \{\mathcal{N}(\mathbf{p}_j'; \mu_j', \Sigma_j')\}_{j=1}^K \tag{6}
\]

such that when aggregated, the selected Gaussians approximate the original MoG distribution \( f \) of Eq. (5), formed by all \( N \) detections. That is, if \( g \) is defined by

\[
g(\mathbf{p}) = \sum_{j=1}^K \beta_j g_j(\mathbf{p}), \tag{7}\]

then we seek a mixture of \( K \) Gaussians, \( G \), for which

\[
d(f, g) = \sum_{i=1}^N \alpha_i \min_{j=1}^K \text{KL}(f_i || g_j), \tag{8}\]
is minimized, where KL is the KL-divergence \([24]\) used as a non-symmetric distance between two detection boxes.

**An EM-approach for selecting detections.** We approximate a solution to minimization of Eq. (8) using an EM-based method. The E-step assigns each box to the nearest box cluster, where box similarity is defined by a KL distance between the corresponding Gaussians. E-step assignments are defined as:

\[
\pi(i) = \arg \min_{j=1}^{K} \text{KL}(f_i || g_j).
\]

The M-step then re-estimates the model parameters by:

\[
\beta_j = \sum_{i \in \pi^{-1}(j)} \alpha_i
\]
\[
\mu'_j = \frac{1}{\beta_j} \sum_{i \in \pi^{-1}(j)} \alpha_i \mu_i
\]
\[
\Sigma'_j = \frac{1}{\beta_j} \sum_{i \in \pi^{-1}(j)} \alpha_i \left( \Sigma_i + (\mu_i - \mu'_j)(\mu_i - \mu'_j)^\top \right).
\]

Note that these matrix computations are fast in 2D space. Moreover, all our Gaussians represent axis-aligned detection and so they all have diagonal covariances. In such cases, the KL distance between two Gaussians has a simpler form which is even more efficient to compute.

General EM theory guarantees that the iterative process described in Eq. (9)-(10), is monotonically decreasing in the value of Eq. (8) and converging to a local minimum \([11]\). We determine convergence when the value of Eq. (8) is smaller than \(\epsilon_{EM} = 1e^{-10}\). We found this process to nearly always converge within ten iterations and so we set a maximum number of iterations at that number.

EM parameters are often initialized using fast clustering to prevent convergence to poor local minima. We initialize it with an agglomerative, hierarchical clustering \([39]\), where each detection initially represents a cluster of its own and clusters are successively merged until \(K\) clusters remain.

We note in passing that there have been several recent attempts to develop deep clustering methods \([47, 48]\). Such methods are designed for clustering high-dimensional data, training autoencoders to map input data into a low-dimensional feature space where clustering is easier. We instead use EM, as these methods are not relevant in our settings, where the original data is two-dimensional.

**Gaussians as detections.** Once EM converged, the estimated Gaussians represent a set of \(K\) detections. As an upper bound for the number of detections, we use \(K = \text{size}(1)/\|\mu_{\text{min}}\|_b\) approximating the amount of non-overlapping, mean-sized boxes that fit into the image. As post-processing, we suppress less confident Gaussians which overlap other Gaussians by more than a predefined threshold. This step can be viewed as model selection and it determines the actual number of detected objects, \(K' \leq K\).

To extract the final detections, for each of the \(K'\) Gaussians, we consider the ellipse at two standard deviations around its center, visualized in Fig. 3 in green. We then search the original set of \(N\) detections (Sec. 3.1) for those whose center, \(\mu = (x, y)\), falls inside this ellipse. A Gaussian is converted to a detection window by taking the median dimensions of the detections in this set.

### 4. The SKU-110K benchmark

We assembled a new labeled data set and benchmark containing images of supermarket shelves. We focus on such retail environments for two main reasons. First, to maximize sales and store real-estate usage, shelves are regularly optimized to present many items in tightly packed, efficient arrangements \([8, 35]\). Our images therefore represent extreme examples of dense environments; precisely the type of scenes we are interested in.

Second, retail items naturally fall into product, brand, and sub-brand object classes. Different brands and products are designed to appear differently. A typical store can sell hundreds of products, thereby presenting a detector with many inter-class appearance variations. Sub-brands, on the other hand, are often distinguishable only by fine-grained packaging differences. These subtle appearance variations increase the range of nuisances that detectors must face (e.g., spatial transformations, image quality, occlusion).

As we show in Table I, SKU-110K is very different from existing alternatives in the numbers and density of the objects appearing in each image, the variability of its item classes, and, of course, the nature of its scenes. Example images from SKU-110K are provided in Fig. 1, 2, and 5.

**Image collection.** SKU-110K images were collected from thousands of supermarket stores around the world, including locations in the United States, Europe, and East Asia. Dozens of paid associates acquired our images, using their personal cellphone cameras. Images were originally taken at no less than five mega-pixel resolution but were then JPEG compressed at one megapixel. Otherwise, phone and camera models were not regulated or documented. Image quality and view settings were also unregulated and so our images represent different scales, viewing angles, lighting conditions, noise levels, and other sources of variability.

Bounding box annotations were provided by skilled annotators. We chose experienced annotators over unskilled, Mechanical Turkers, as we found the boxes obtained this way were more accurate and did not require voting schemes to verify correct annotations \([28, 42]\). We did, however, visually inspect each image along with its detection labels, to filter obvious localization errors.

**Benchmark protocols.** SKU-110K images were parti-
**Table 2. Detection runtime comparison on SKU-110K.**

<table>
<thead>
<tr>
<th>Method</th>
<th>FPS</th>
<th>DPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN (2015) [38]</td>
<td>2.37</td>
<td>93</td>
</tr>
<tr>
<td>YOLO9000 (2017) [37]</td>
<td>5</td>
<td>317</td>
</tr>
<tr>
<td>RetinaNet (2018) [27]</td>
<td>0.5</td>
<td>162</td>
</tr>
<tr>
<td>Base detector</td>
<td>0.5</td>
<td>162</td>
</tr>
<tr>
<td>+ Soft-IoU</td>
<td>0.23</td>
<td>73</td>
</tr>
<tr>
<td>+ EM-Merger (on the CPU)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. Detection on SKU-110K. Bold numbers are best results.**

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP 75</th>
<th>AR 300</th>
<th>P R=0.5</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monkey</td>
<td>.000</td>
<td>.010</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faster-RCNN [38]</td>
<td>.045</td>
<td>.010</td>
<td>.066</td>
<td>0</td>
<td>107.46</td>
<td>113.42</td>
</tr>
<tr>
<td>YOLO9000 opt [37]</td>
<td>.094</td>
<td>.073</td>
<td>.111</td>
<td>0</td>
<td>84.166</td>
<td>97.809</td>
</tr>
<tr>
<td>RetinaNet [27]</td>
<td>.455</td>
<td>.389</td>
<td>.530</td>
<td>.544</td>
<td>16.584</td>
<td>30.702</td>
</tr>
<tr>
<td>Base &amp; NMS</td>
<td>.413</td>
<td>.384</td>
<td>.484</td>
<td>.491</td>
<td>24.962</td>
<td>34.382</td>
</tr>
<tr>
<td>Soft-IoU &amp; NMS</td>
<td>.418</td>
<td>.386</td>
<td>.483</td>
<td>.492</td>
<td>25.394</td>
<td>34.729</td>
</tr>
<tr>
<td>Base &amp; EM-Merger</td>
<td>.482</td>
<td>.540</td>
<td>.553</td>
<td>.802</td>
<td>23.978</td>
<td>283.971</td>
</tr>
<tr>
<td>Our full approach</td>
<td>.492</td>
<td>.556</td>
<td>.554</td>
<td>.834</td>
<td>14.522</td>
<td>23.992</td>
</tr>
</tbody>
</table>

**Evaluation.** We adopt evaluation metrics similar to those used by COCO [28], reporting the average precision (AP) at IoU=.50:.05:.95 (their primary challenge metric), AP at IoU=.75, AP\(^75\) (their strict metric), and average recall (AR\(^300\)) at IoU=.50:.05:.95 (300 is the maximal number of objects). We further report the value sampled from the precision-recall curve at recall = 0.5 for IoU=0.75 (P\(_R\)=0.5).

The many, densely packed items in our images are reminiscent of the settings in counting benchmarks [2, 22]. We capture both detection and counting accuracy, by borrowing the error measures used for those tasks: If \(\{K'_i\}_{i=1}^n\) is the predicted numbers of objects in each test image, \(i \in [1, n]\), and \(\{t_i\}_{i=1}^n\) are the per image ground truth numbers, then the mean absolute error (MAE) is \(\frac{1}{n} \sum_{i=1}^{n} |K'_i - t_i|\) and the root mean squared error (RMSE) is \(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (K'_i - t_i)^2}\).

**5. Experiments**

**5.1. Run-time analysis**

Table 2 compares average frames per second (FPS) and detections per second (DPS) for baseline methods and variations of our approach. Runtimes were measured on the same machine using an Intel(R) Core(TM) i7-5930K CPU @3.50GHz GeForce and a GTX Titan X GPU.

Our base detector is modeled after RetinaNet [27] and so their runtimes are identical. Adding our Soft-IoU layer does not affect runtime. EM-Merger is slower despite the optimizations described in Sec. 3.3 mostly because of memory swapping between GPU and CPU/RAM. Our initial tests suggest that a GPU optimized version will be nearly as fast as the base detector.

**5.2. Experiments on the SKU-110K benchmark**

**Baseline methods.** We compare the detection accuracies of our proposed method and recent state-of-the-art on the SKU-110K benchmark. All methods, with the exception of the Monkey detector, were trained on the training set portion SKU-110K.

The following two baseline methods were tested using the original implementations released by their authors: RetinaNet [27] and Faster-RCNN [38]. YOLO9000 [37] is not suited for images with more than 50 objects. We offer results for YOLO9000 opt, which is YOLO9000 with its loss function optimized and retrained to support detection of up to 300 boxes per image.

We also report the following ablation studies, detailing the contributions of individual components of our approach.

- **Monkey:** Because of the tightly packed items in SKU-110K images, it is plausible that randomly tossed bounding boxes would correctly predict detections by chance. To test this naive approach, we assume we know the object number, \(K'\), the mean and standard-
deviation width, $\mu_w$, $\sigma_w$, and height, $\mu_h$, $\sigma_h$, for these boxes. Monkey samples 2D upper-left corners for the $K'$ bounding boxes from a uniform distribution and box heights and widths from Gaussian distributions $\mathcal{N}(h; \mu_h, \sigma_h)$ and $\mathcal{N}(h; \mu_w, \sigma_w)$, respectively.

- **Base & NMS**: Our basic detector of Sec. 3.1 with standard NMS applied to objectness scores, $c$.
- **Soft-IoU & NMS**: Base detector with Soft-IoU (Sec. 3.2). Standard NMS applied to Soft-IoU scores, $c_{iou}$, instead of objectness scores.
- **Base & EM-Merger**: Our basic detector, now using EM-Merger of Sec. 3.3 but applying it to original objectness scores, $c$.
- **Our full approach**: Applying the EM-Merger unit to Deep-IoU scores, $c_{iou}$.

To test MAE and RMSE we report the number of detected objects, $K'$, and compare it with the true number of items per image. In RetinaNet the number of detections is extremely high so we first filter detections with low confidences. This confidence threshold was determined using cross-validation to optimize the results of this baseline.

**Detection results on SKU-110K**. Quantitative detection results are provided in Table 3, result curves are presented in Fig. 4, and a selection of qualitative results, comparing

![Figure 5. Qualitative detection results on SKU-110K. Please see project web-page for more results and images in higher resolutions.](image-url)
Finally, we note that our best results remain far from perfect: The densely packed settings represented by SKU-110K images appear to be highly challenging, leaving room for further improvement.

5.3. Experiments on CARPK and PUCPR+

We test our method on data from other benchmarks, to see if our approach generalizes well to other domains beyond store shelves and retail objects. To this end, we use the recent CARPK and PUCPR+ benchmarks. Both data sets provide images of parking lots from high vantage points. We use their test protocols, comparing the number of detections per image to the ground truth numbers made available by these benchmarks. Accuracy is reported using MAE and RMSE, as in our SKU-110K (Sec. 4).

**Counting results.** We compare our method with results reported by others [22, 41]: Faster R-CNN [38], YOLO [36], and One-Look Regression [32]. Existing baselines also include two methods designed and tested for counting on these two benchmarks: LPN Counting [22] and IEP Counting [41]. In addition, we trained and tested counting accuracy with YOLO9000opt [37] and RetinaNet [27].

Table 4 reports the MAE and RMSE for all tested methods. Despite not being designed for counting, our method is more accurate than recent methods designed for that task. A significant difference between these counting datasets and our SKU-110K is in the much closer proximity of the objects in our images. This issue has a significant impact on baseline detectors, as can be seen in Tables 4 and 5. Our model suffers a much lower degradation in performance due to better filtering of these overlaps.

6. Conclusions

The performance of modern object/no-object detectors on existing benchmarks is remarkable yet still limited. We focus on densely packed scenes typical of every-day retail environments and offer SKU-110K, a new benchmark of such retail shelf images, labeled with item detection boxes. Our tests on this benchmark show that such images challenge state-of-the-art detectors.

To address these challenges, along with our benchmark, we offer two technical innovations designed to raise detection accuracy in such settings: The first is a Soft-IoU layer for estimating the overlap between predicted and (unknown) ground truth boxes. The second is an EM-based unit for resolving bounding box overlap ambiguities, even in tightly packed scenes where these overlaps are common.

We test our approach on SKU-110K and two existing benchmarks for counting, and show it to surpass existing detection and counting methods. Still, even the best results on SKU-110K are far from saturated, suggesting that these densely packed scenes remain a challenging frontier for future work.

---

1See project web-page for qualitative results on these benchmarks.
References


