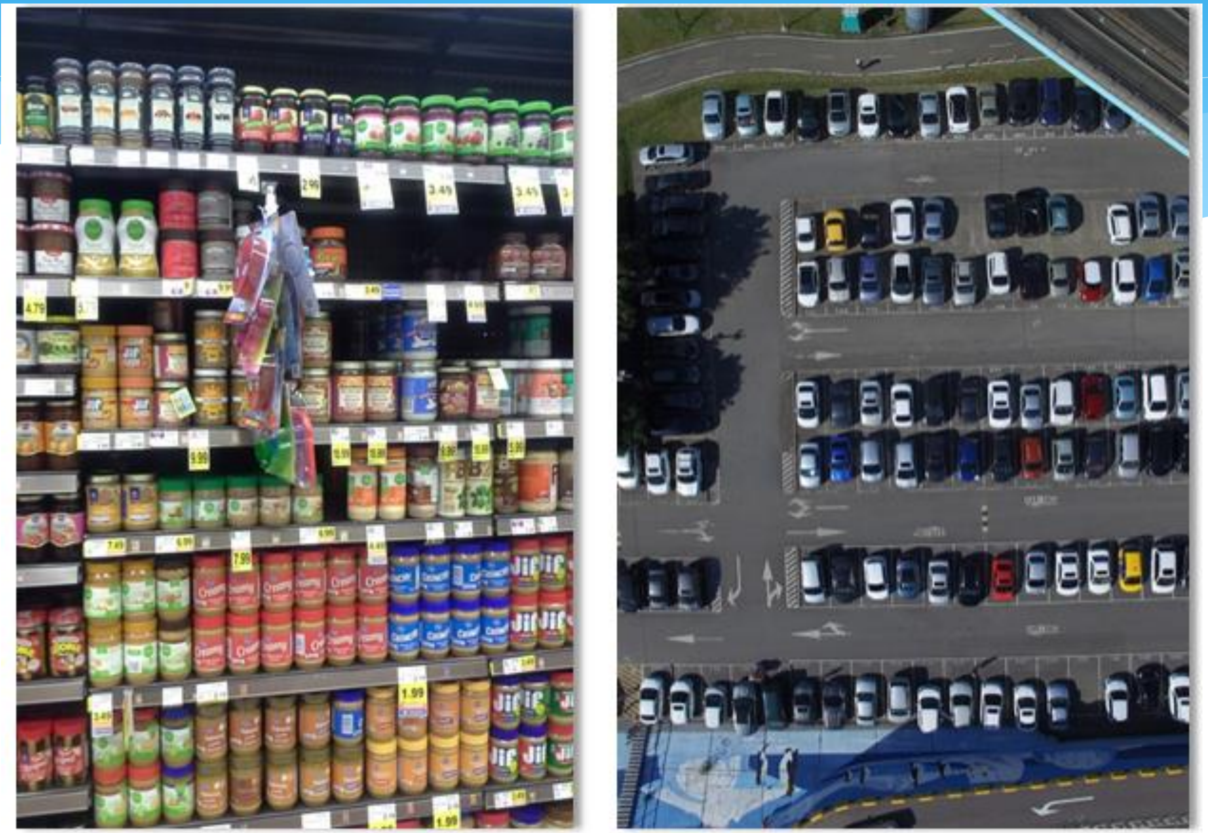
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION

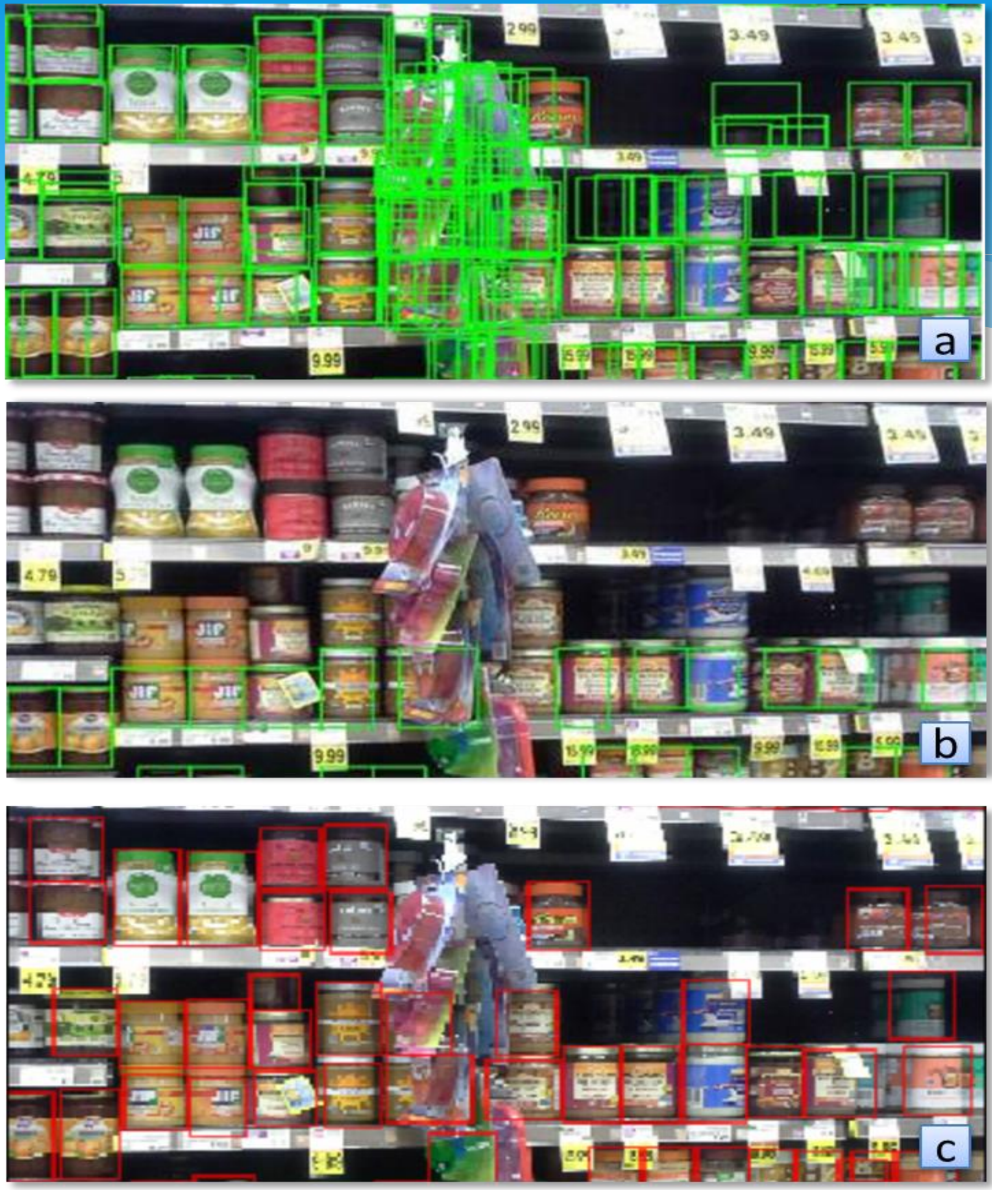
Multiple-step Sampling for Dense Object Detection and Counting

Zhaoli Deng, Chenhui Yang*
School of Informatics
Xiamen University
Xiamen, China
Email: dzleee@stu.xmu.edu.cn, chyang@xmu.edu.cn

Dense scenes

Dense scenes are common sights in the metropolis, parking lot, retail shelf displays, and landscape images being the prominent ones. The images of dense scenes contain a multitude of similar or even identical objects, which positioned closely, pose a big challenge to the accuracy and efficiency of the detection.





Detection in packed domains.
A typical image in SKU-110K,

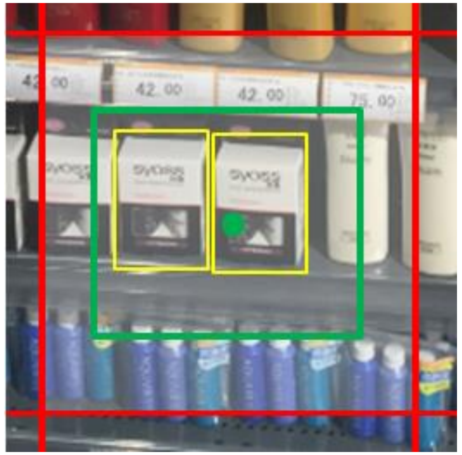
(a) Detection results for the RetinaNet , it returns a lot of overlapping or incorrect boxes
.

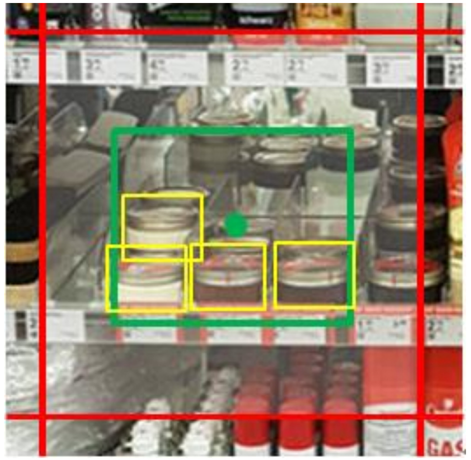
(b) The famous two-stage detector Faster R-CNN detects few objects.

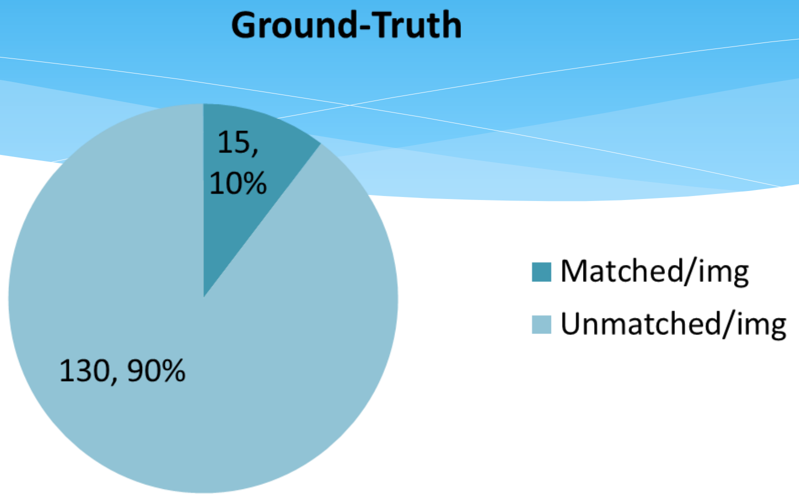
(c) Our approach does precise detection.

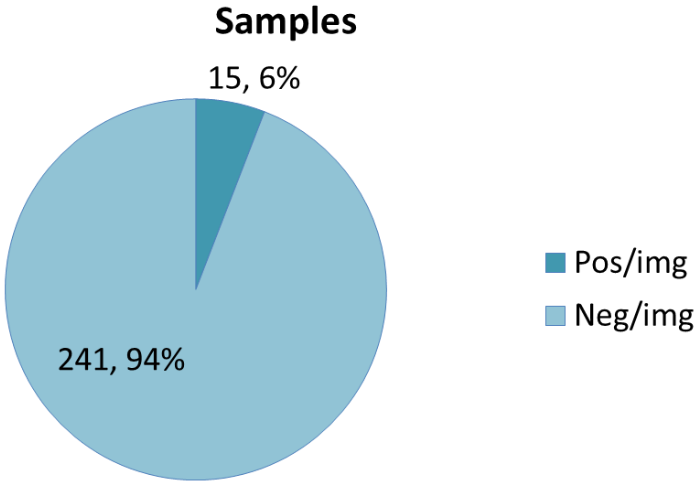
Waste of Ground-truth

In the left images, an anchor may contain a high IoU overlap with two boxes. While in the right, there many small ground-truth boxes in an anchor.









Multiple-step sampling

There are there are N anchors and M ground-truth boxes in an image and we denote the set of anchors as $\{a_i\}_{i=1}^{N}$ and the set of ground-truth boxes as $\{g_j\}_{j=1}^{M}$.

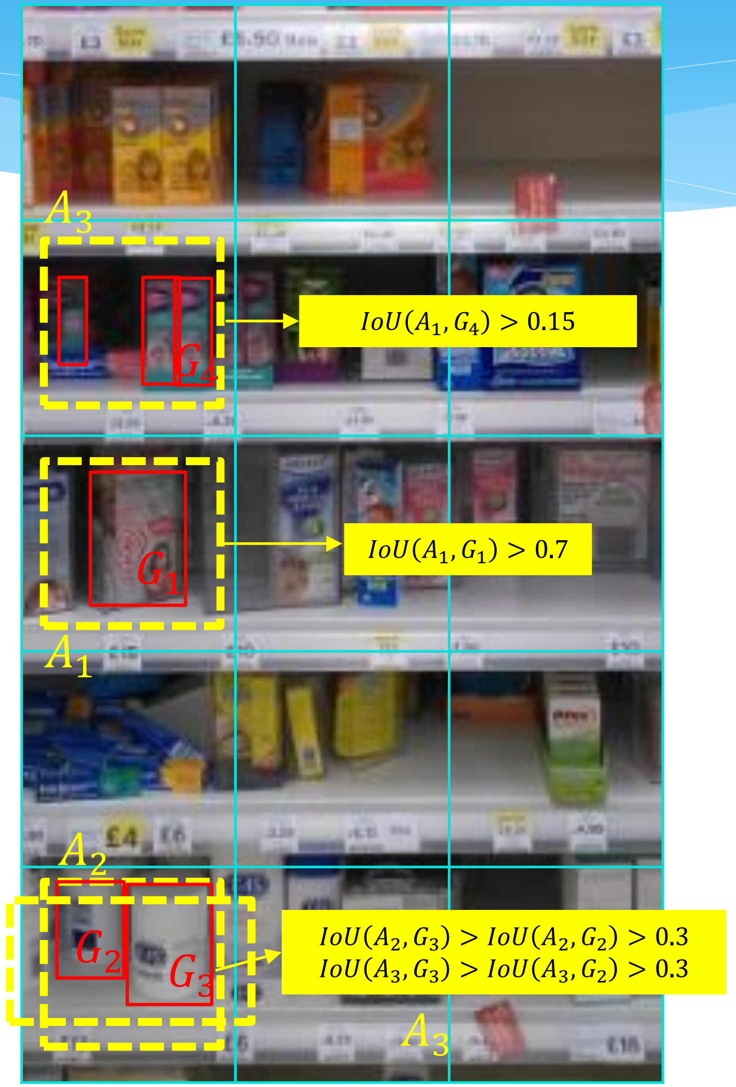
We assign anchors with positive labels based on their IoU ratio in the following three step. In the first step, we associate ground-truth boxes g_j with anchor a_i . They are satisfy:

$$\mathrm{IoU}\left(a_i, g_j\right) \geq 0.7.$$

In the second, we select an anchor a_i without label to match a ground-truth box g_j.

$$\mathrm{IoU}\left(a_i, g_j\right) \geq 0.3.$$

In the third step, we select anchor to match the small groud-truth.

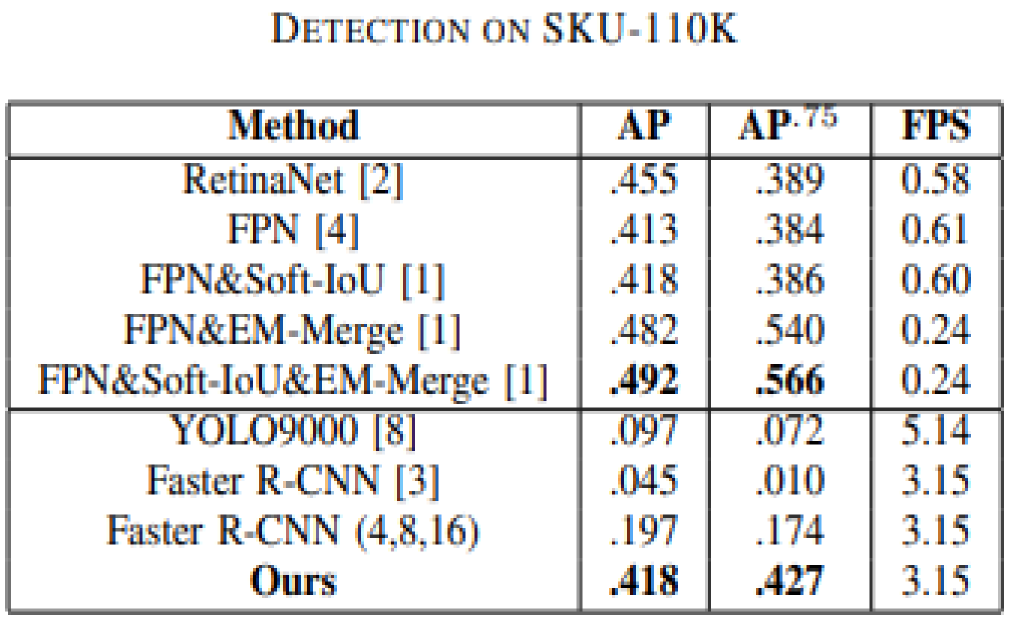


$IoU(A_1, G_4) > 0.15$

$IoU(A_1, G_1) > 0.7$

$IoU(A_2, G_3) > IoU(A_2, G_2) > 0.3$
$IoU(A_3, G_3) > IoU(A_3, G_2) > 0.3$

Experiments



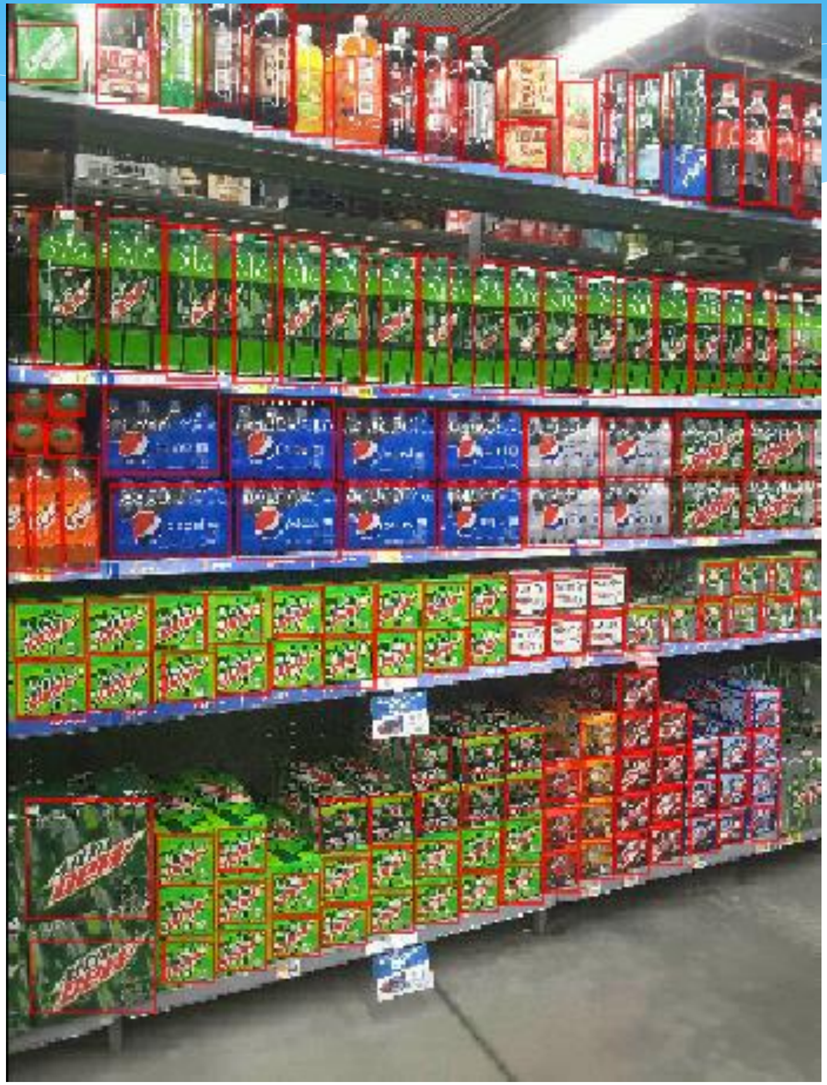
DETECTION ON SKU-110K

Method	AP	AP$^{.75}$	FPS
RetinaNet [2]	.455	.389	0.58
FPN [4]	.413	.384	0.61
FPN&Soft-IoU [1]	.418	.386	0.60
FPN&EM-Merge [1]	.482	.540	0.24
FPN&Soft-IoU&EM-Merge [1]	**.492**	**.566**	0.24
YOLO9000 [8]	.097	.072	5.14
Faster R-CNN [3]	.045	.010	3.15
Faster R-CNN (4,8,16)	.197	.174	3.15
Ours	**.418**	**.427**	3.15

We divide results into two groups. The first group is detectors with FPS below 1.0, while the second is higher than 1.0. Our full approach with the best detection accuracy in the second group and a lot ahead of other detectors.
Compare our method with the first group, our approach has higher AP (.418) than FPN's (.413) and higher AP.75 (.427) than RetinaNet (.389). In contrast with the detectors with the best accuracy, we can predict most bounding boxes without overlap. Moreover, it is noteworthy that our FPS (3.15fps) is more than five times that of RetinaNet (0.58fps) and thirteen times that of FPN&Em-merge (0.24fps), claiming a leading position in detection speed.

Conclusions

It is observed that training samples imbalance causes Faster R-CNN to perform worse. While this imbalance is the result of the waste of ground-truth boxes. Motivated by this, we propose a multiple-step sampling method, featuring in simple procedures and effective outcomes, to balance the training samples. We test our method on SKU-110K and CARPK benchmarks. Our method can be applied to Faster R-CNN to further increase detection accuracy. The improved Faster R-CNN performs as well as the state-of-the-art with higher inference speed.



Thanks!