



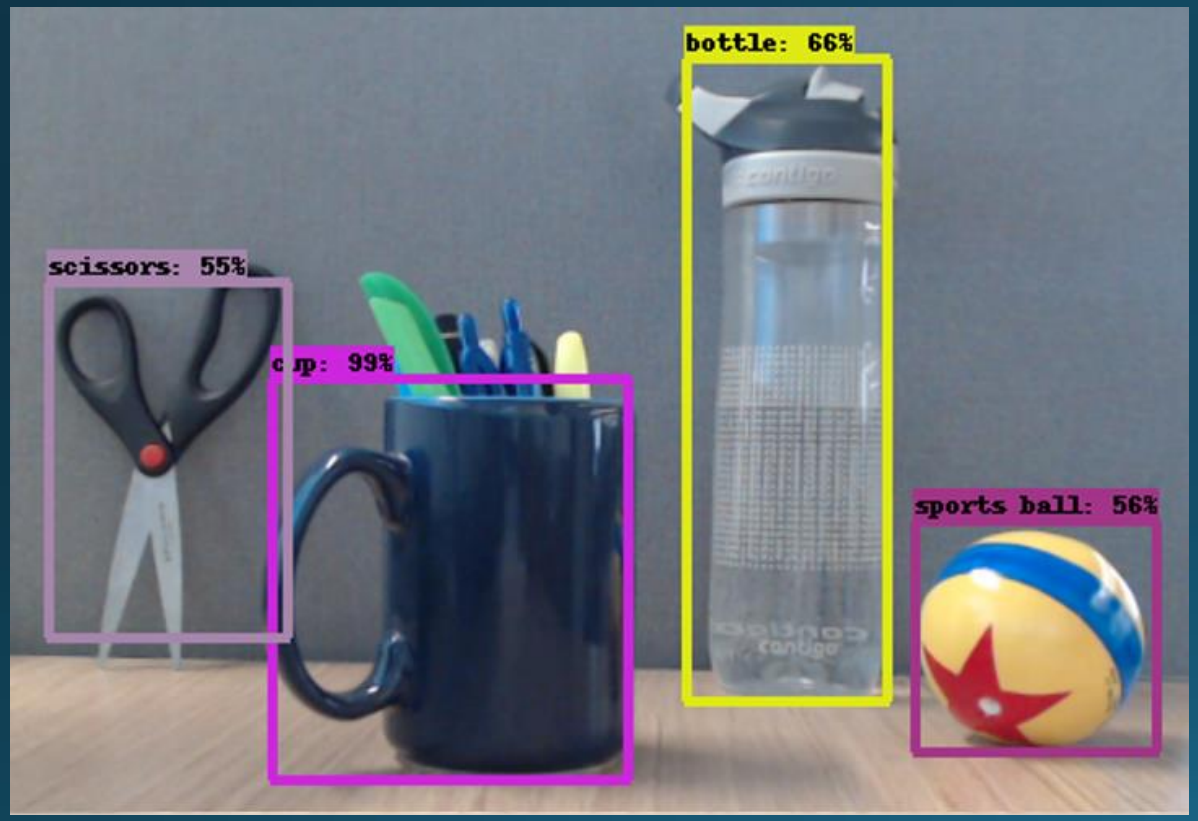
# Precise Detection in Densely Packed Scenes

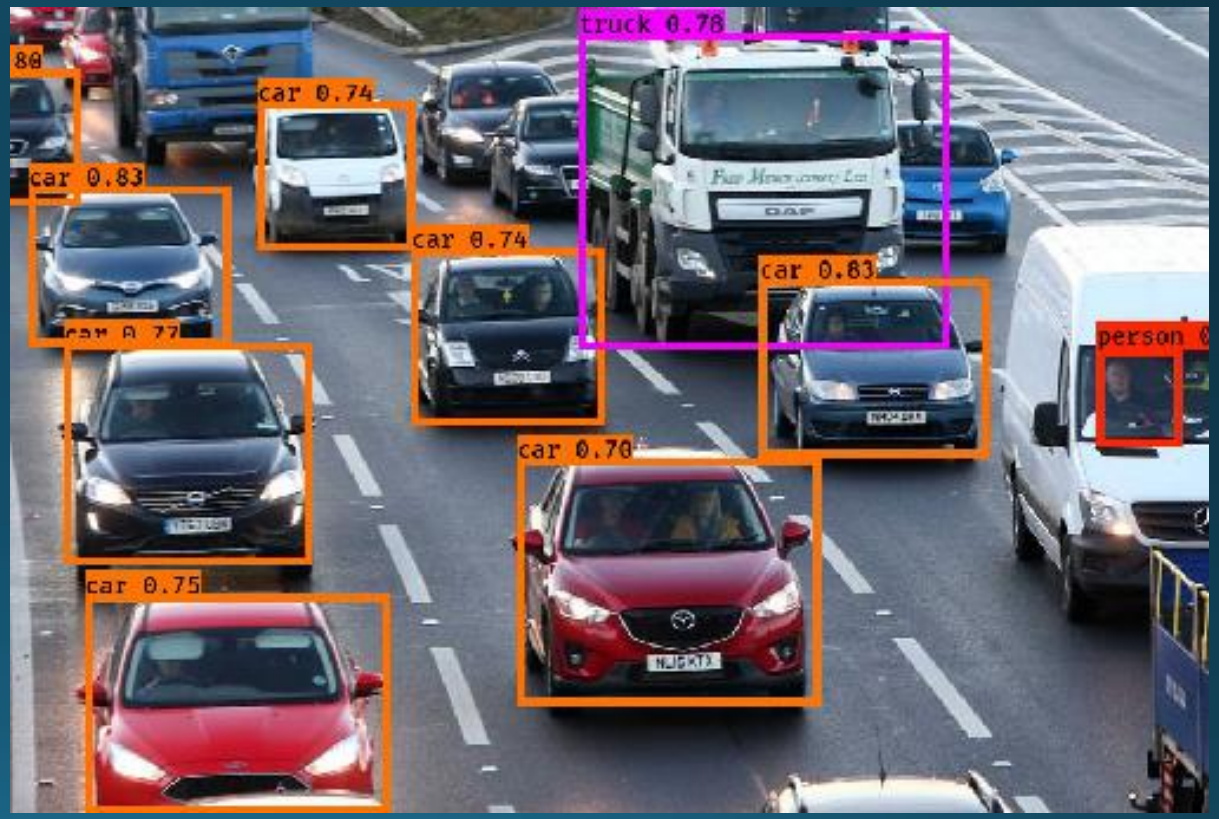
*Eran Goldman*[1,2] *Roei Herzig*[4] *Aviv Eisenschtat*[1] *Oria Ratzon*[1] *Itsik Levi*[1]

*Jacob Goldberger*[2] *Tal Hassner*[3]

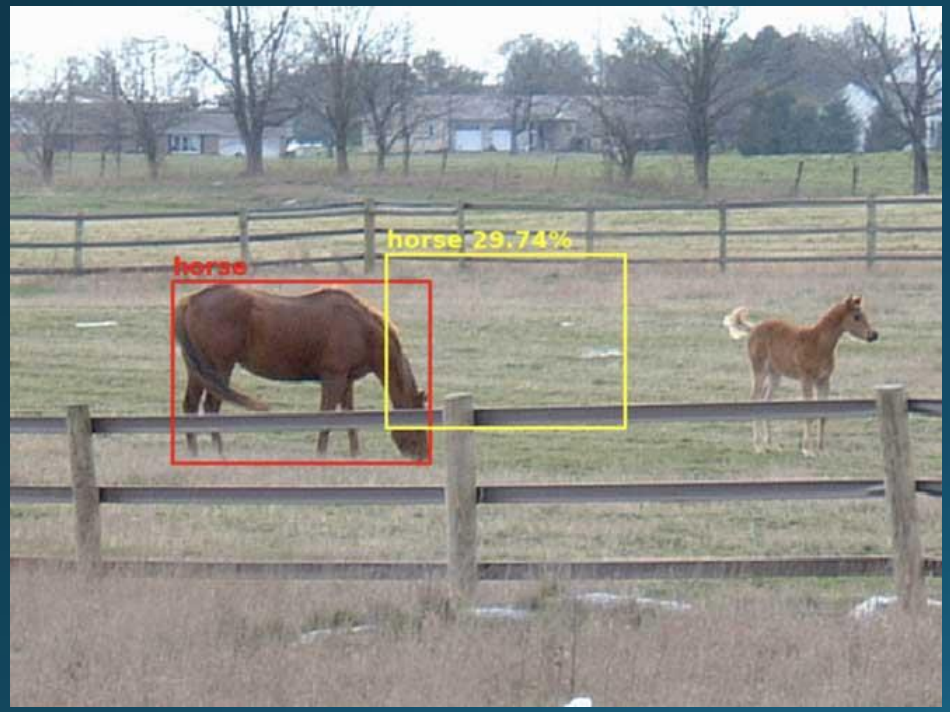
**1. Trax Retail, 2. Bar-Ilan University , 3. Open University of Israel, 4. Tel Aviv University**

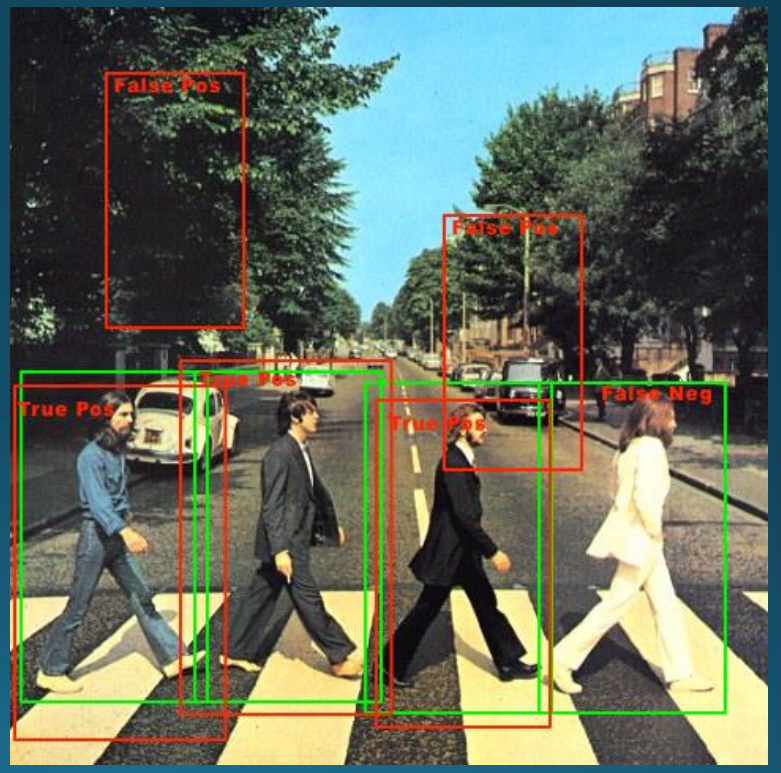
# Object Detection



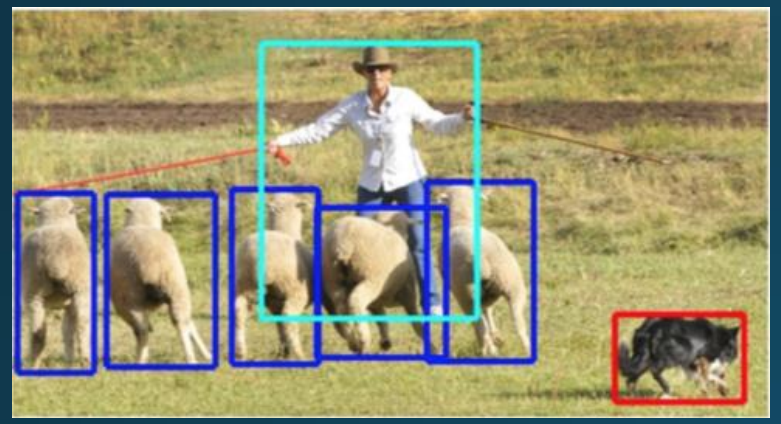



# False Detections

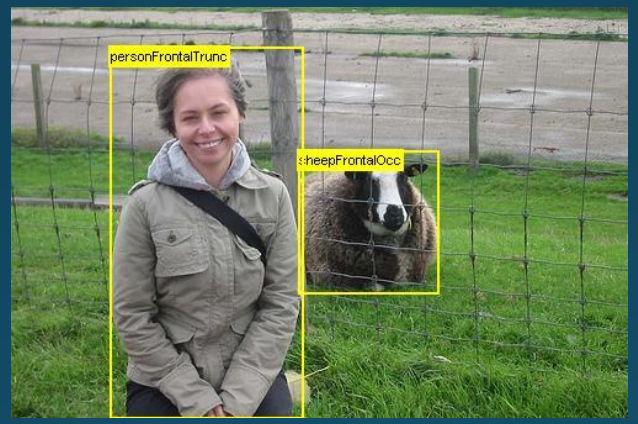


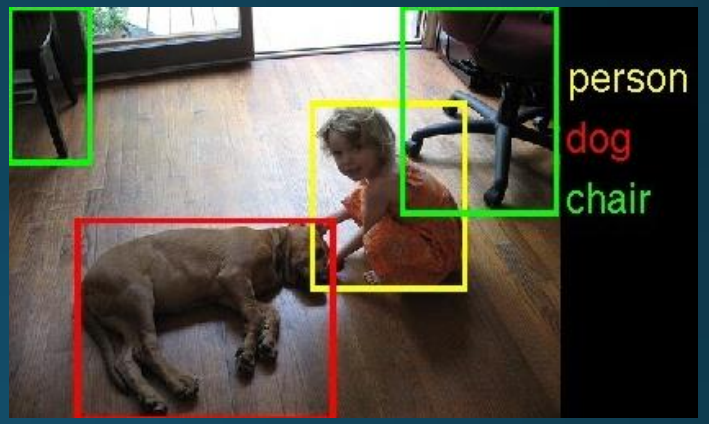
# Natural Images Datasets



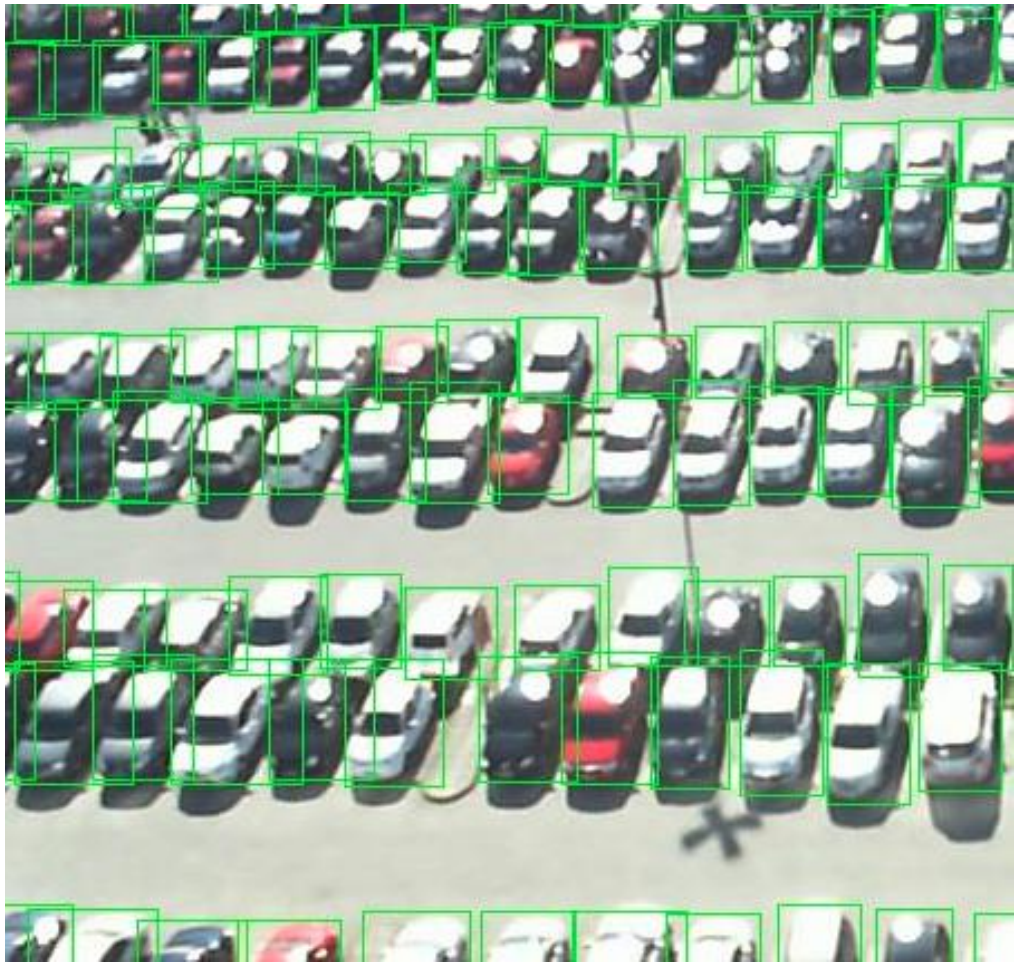
COCO




VOC




ImageNet

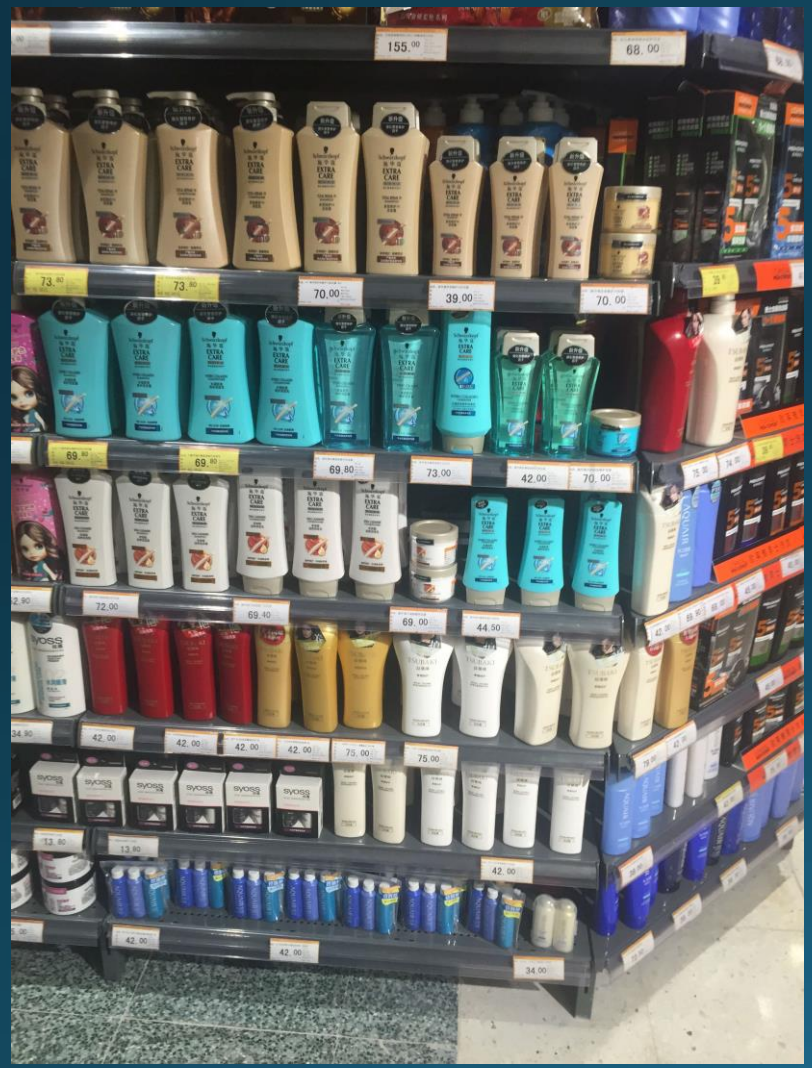


10th floor view

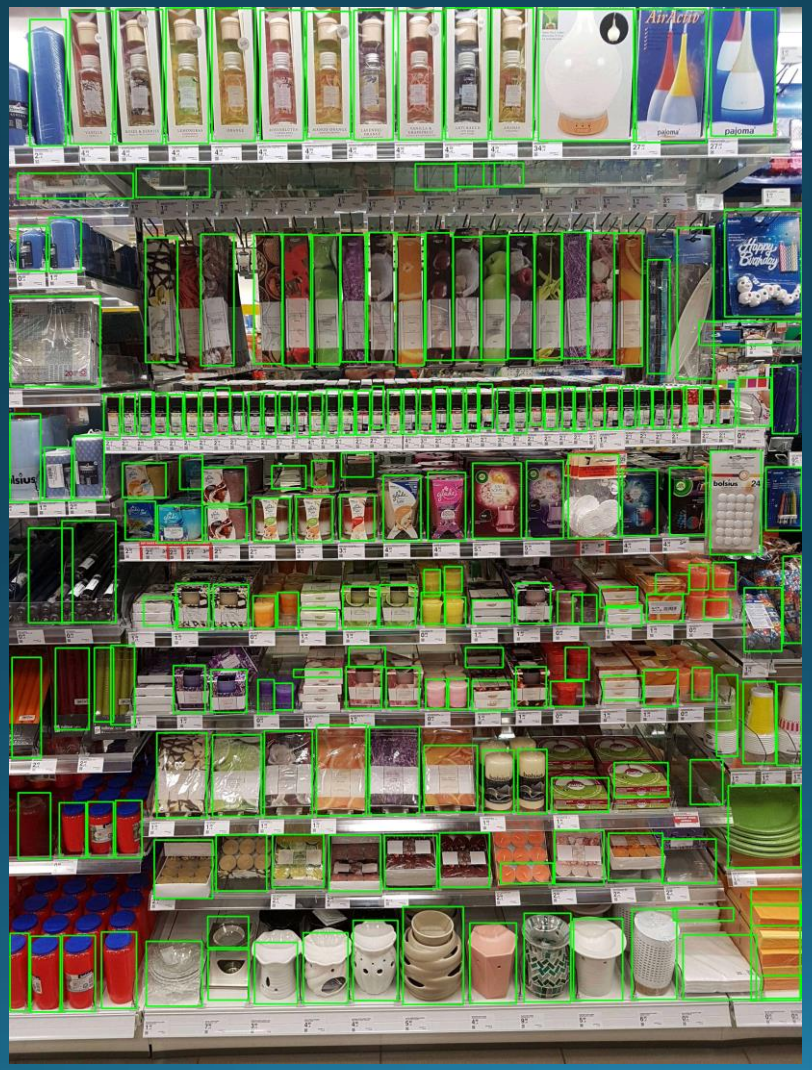
Drone view

# Object Detection in Retail

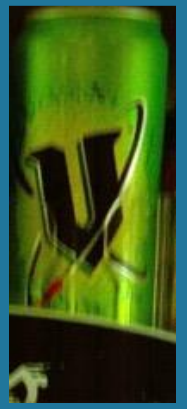
trax

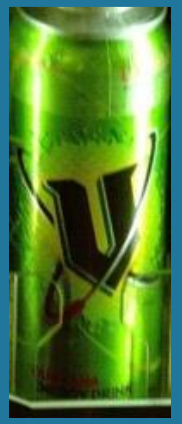


# Object Detection in Retail



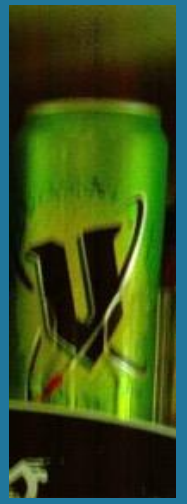
# Are these accurate detections of  cans?



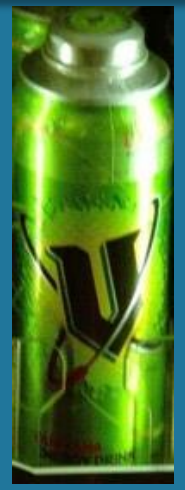


# Are these accurate detections of  cans?

Yes!



Not Really



No!

Really Not

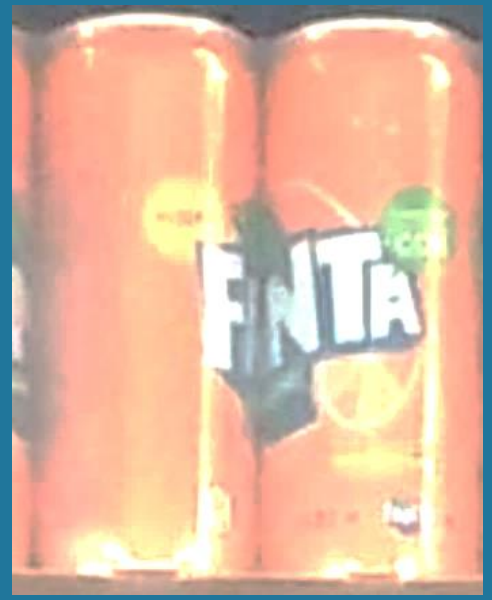
# How many objects are here?


# How many objects are here?

1

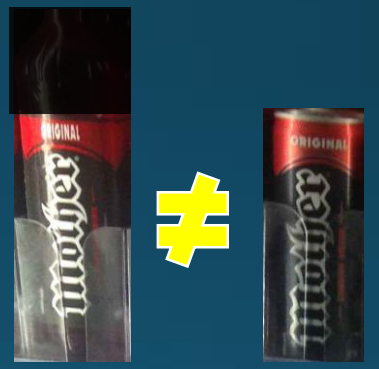

4

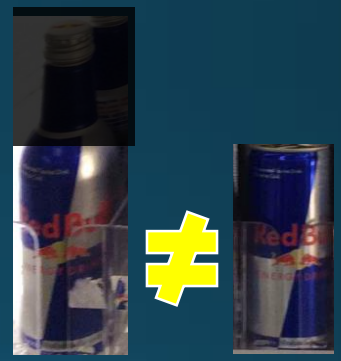

2



# Key insights:

## Appearance ≠ Localization accuracy



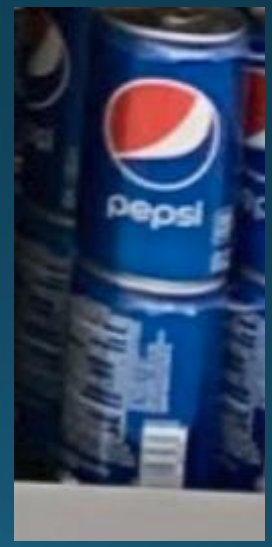

# Key insights:

Appearance ≠ Localization accuracy

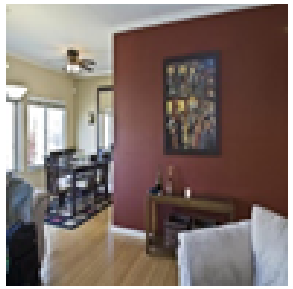

# Key insights:

Appearance ≠ Localization accuracy

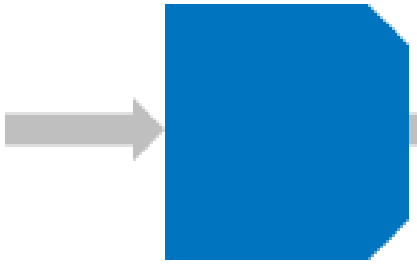




# Object Detector Architecture



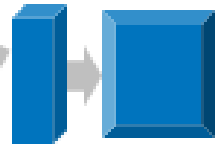
Images



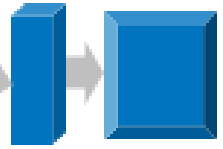
Multiple Layers of Convolution and Pooling

Feature Map

Classification Head

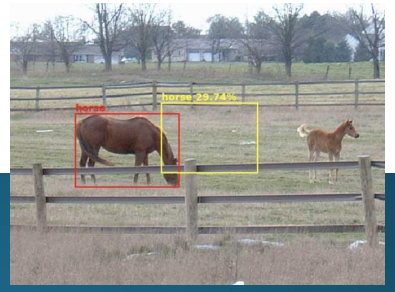


Object Class

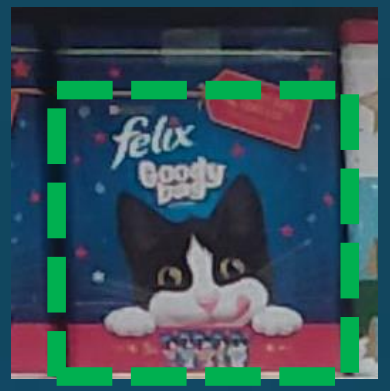


Box Coordinates (x, y, width, height)

Regression Head

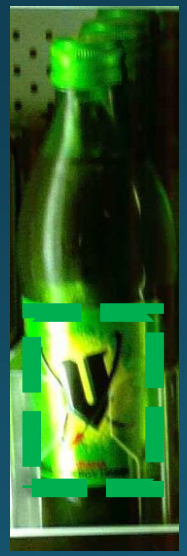


Object

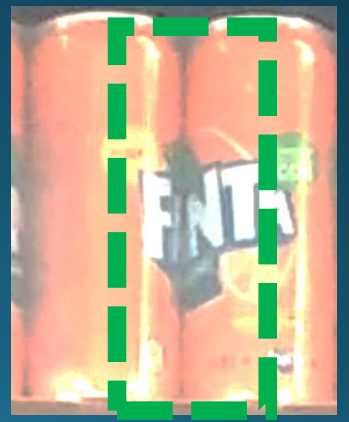


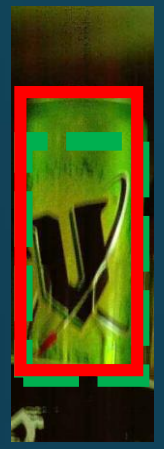
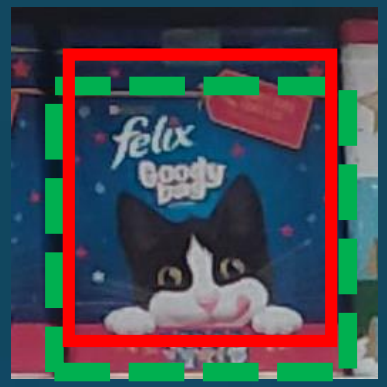


Non Object

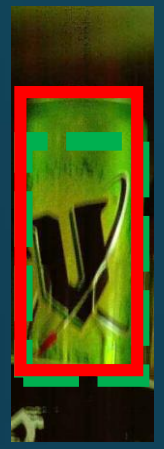


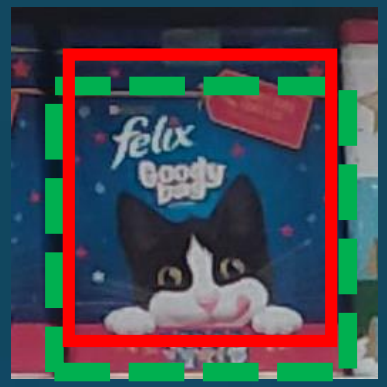




| | | | |
|---|---|---|---|
| Object |  |  |  |
| Non Object |   |  |  |

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

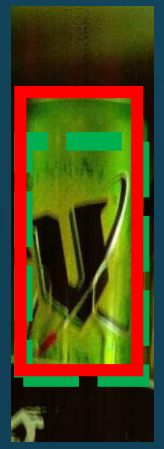




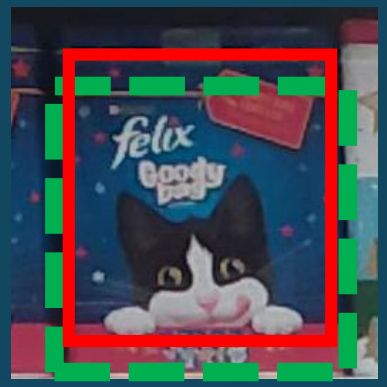




$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



0.8




0.83


0.94

0.3

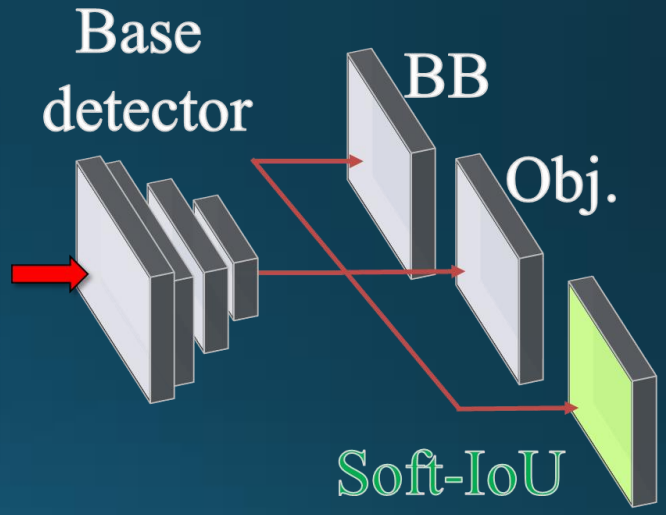
0.17


0.37


0.45

# Novel IoU layer

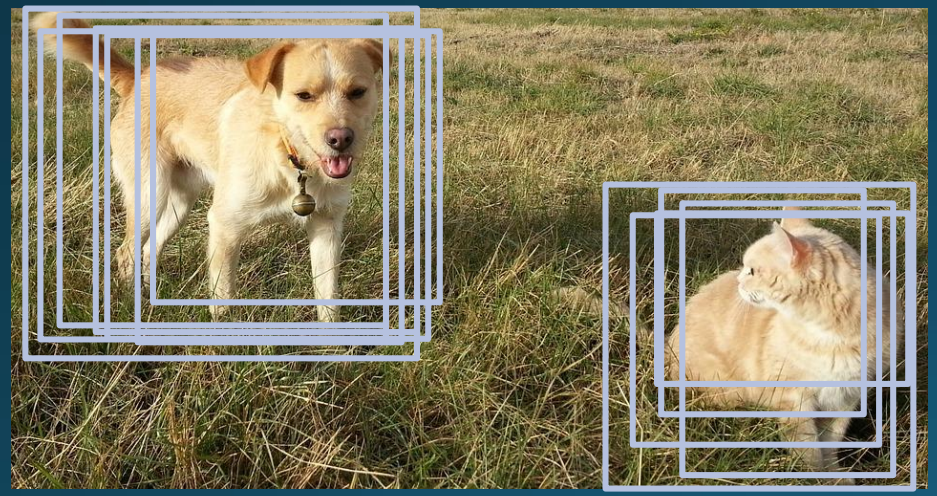
- Predict IoU rates between detected and true boxes
- Take a probabilistic interpretation of the IoU rate function and learn it with cross-entropy loss




$$\mathcal{L}_{sIoU} = -\frac{1}{n}\sum_{i=1}^{n}[IoU_i \log(c_i^{iou}) + (1 - IoU_i)\log(1 - c_i^{iou})]$$

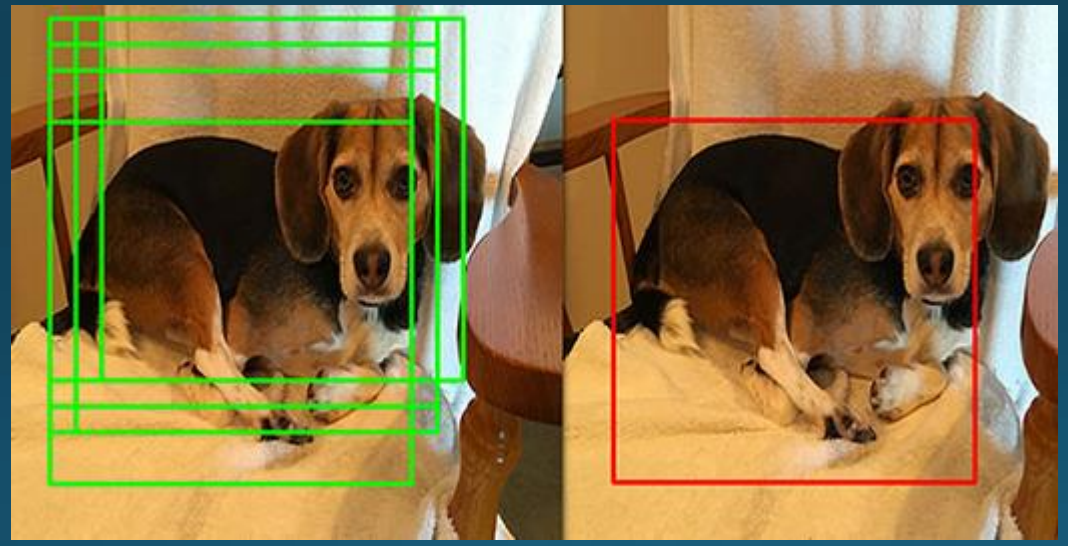
# Resolving Duplicates

The "Standard" Case

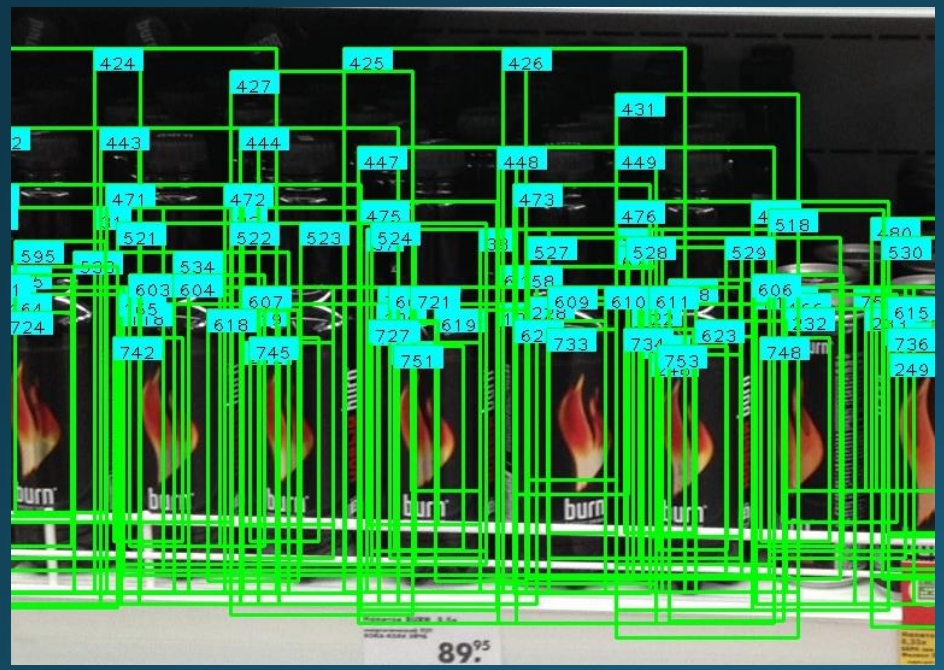


Our Case

# Non-maximum suppression (NMS)



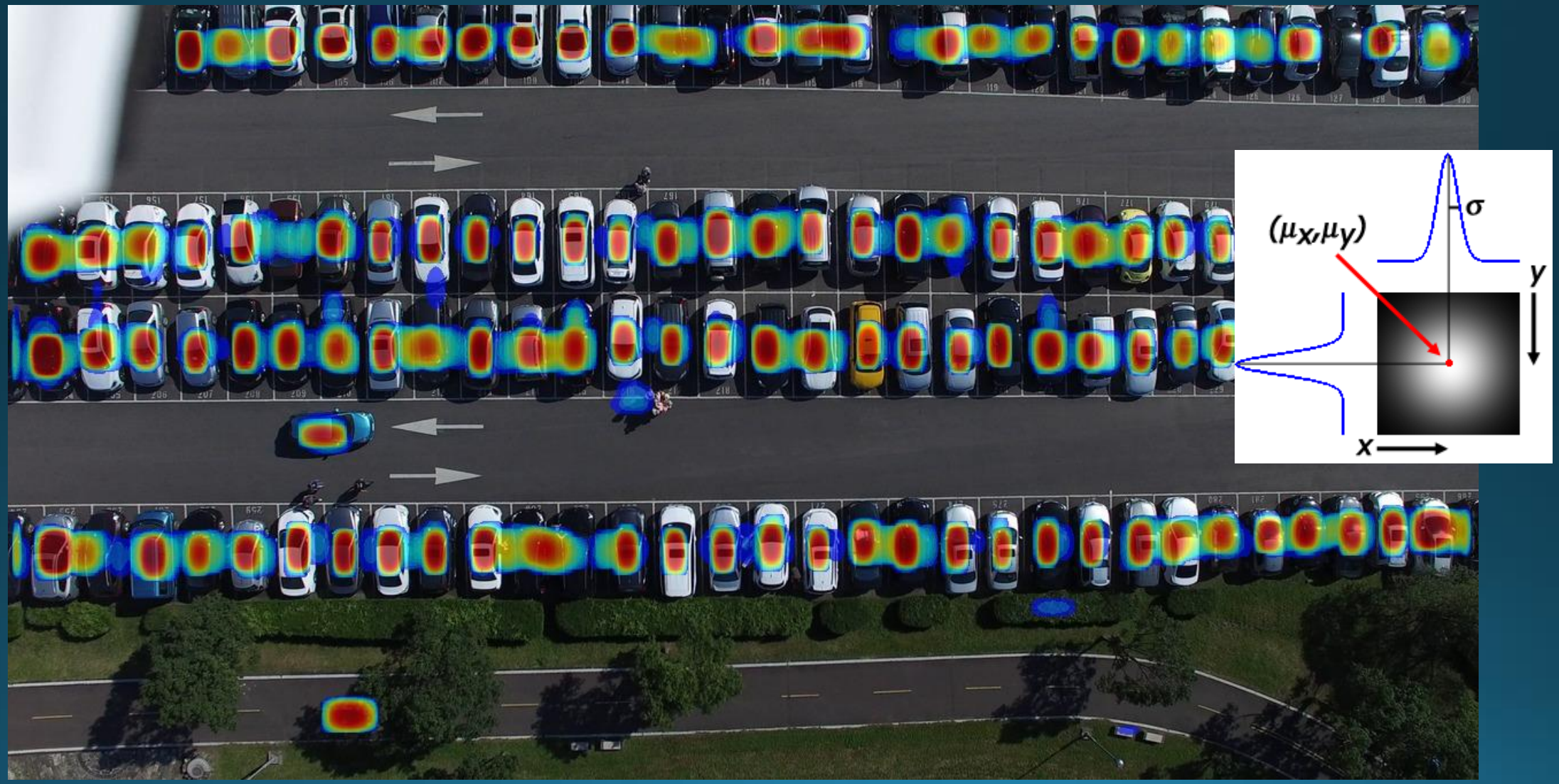
# Non-maximum suppression (NMS)




# Resolve overlapping detections

# Resolve overlapping detections

# Resolve overlapping detections



# GMM Many-to-Few EM

# EM-Merger unit

Resolve overlapping detections

- Model original detections as a GMM
- Use EM to reduce number of GMM components
- Convert the reduced GMM to final detections

E-step: $\pi(t) = \min_{i \le 1 j \le k} KL(f_t || g_j)$

M-step: $\beta_j = \sum_{t \in \pi^{-1}(j)} \alpha_t$

$$\mu'_j = \frac{1}{\beta_j} \sum_{t \in \pi^{-1}(j)} \alpha_t \mu_t$$

$$\Sigma'_j = \frac{1}{\beta_j} \sum_{t \in \pi^{-1}(j)} \alpha_t \left(\Sigma_t + \left(\mu_t - \mu'_j\right)\left(\mu_t - \mu'_j\right)^T\right)$$


- ✓ Agglomerative initialization
- ✓ 2D matrix computations
- ✓ Fast convergence

trax

- 1,000,000+ objects
- 200+ of objects per image
- Huge variation in items and appearances
- Structured, Crowded, Fine-grained

# SKU-110K dataset

| Name | #Img. | #Obj./img. | #Cls. | #Cls./img. | Dense. | Idnt. | BB |
|---|---|---|---|---|---|---|---|
| UCSD (2008) | 2000 | 24.9 | 1 | 1 | ✓ | ✗ | ✗ |
| PACAL VOC (2012) | 22,531 | 2.71 | 20 | 2 | ✗ | ✗ | ✓ |
| ILSVRC Detection (2014) | 516,840 | 1.12 | 200 | 2 | ✗ | ✗ | ✓ |
| COCO (2015) | 328,000 | 7.7 | 91 | 3.5 | ✗ | ✗ | ✓ |
| Penguins (2016) | 82,000 | 25 | 1 | 1 | ✓ | ✗ | ✗ |
| TRANCOS (2016) | 1,244 | 37.61 | 1 | 1 | ✓ | ✓ | ✗ |
| WIDER FACE (2016) | 32,203 | 12 | 1 | 1 | ✗ | ✗ | ✓ |
| CityPersons (2017) | 5000 | 6 | 1 | 1 | ✗ | ✗ | ✓ |
| PUCPR+ (2017) | 125 | 135 | 1 | 1 | ✓ | ✓ | ✓ |
| CARPK (2018) | 1448 | 61 | 1 | 1 | ✓ | ✓ | ✓ |
| Open Images V4 (2018) | **1,910,098** | 8.4 | 600 | 2.3 | ✗ | ✓ | ✓ |
| **Our SKU-110K** | 11,762 | **147.4** | **110,712** | **86** | ✓ | ✓ | ✓ |

tinyurl.com/sku110k

- 1,000,000+ objects
- 200+ of objects per image
- Huge variation in items and appearances
- Structured, Crowded, Fine-grained

| Name | #Img. | #Obj./img. | #Cls. | #Cls./img. | Dense. | Idnt. | BB |
|---|---|---|---|---|---|---|---|
| UCSD (2008) | 2000 | 24.9 | 1 | 1 | ✓ | ✗ | ✗ |
| PACAL VOC (2012) | 22,531 | 2.71 | 20 | 2 | ✗ | ✗ | ✓ |
| ILSVRC Detection (2014) | 516,840 | 1.12 | 200 | 2 | ✗ | ✗ | ✓ |
| COCO (2015) | 328,000 | 7.7 | 91 | 3.5 | ✗ | ✗ | ✓ |
| Penguins (2016) | 82,000 | 25 | 1 | 1 | ✓ | ✗ | ✗ |
| TRANCOS (2016) | 1,244 | 37.61 | 1 | 1 | ✓ | ✓ | ✗ |
| WIDER FACE (2016) | 32,203 | 12 | 1 | 1 | ✗ | ✗ | ✓ |
| CityPersons (2017) | 5000 | 6 | 1 | 1 | ✗ | ✗ | ✓ |
| PUCPR+ (2017) | 125 | 135 | 1 | 1 | ✓ | ✓ | ✓ |
| CARPK (2018) | 1448 | 61 | 1 | 1 | ✓ | ✓ | ✓ |
| Open Images V4 (2018) | **1,910,098** | 8.4 | 600 | 2.3 | ✗ | ✓ | ✓ |
| **Our SKU-110K** | 11,762 | **147.4** | **110,712** | **86** | ✓ | ✓ | ✓ |

tinyurl.com/sku110k

trax

- 1,000,000+ objects
- 200+ of objects per image
- Huge variation in items and appearances
- Structured, Crowded, Fine-grained

| Name | #Img. | #Obj./img. | #Cls. | #Cls./img. | Dense. | Idnt. | BB |
|---|---|---|---|---|---|---|---|
| UCSD (2008) | 2000 | 24.9 | 1 | 1 | ✓ | ✗ | ✗ |
| PACAL VOC (2012) | 22,531 | 2.71 | 20 | 2 | ✗ | ✗ | ✓ |
| ILSVRC Detection (2014) | 516,840 | 1.12 | 200 | 2 | ✗ | ✗ | ✓ |
| COCO (2015) | 328,000 | 7.7 | 91 | 3.5 | ✗ | ✗ | ✓ |
| Penguins (2016) | 82,000 | 25 | 1 | 1 | ✓ | ✗ | ✗ |
| TRANCOS (2016) | 1,244 | 37.61 | 1 | 1 | ✓ | ✓ | ✗ |
| WIDER FACE (2016) | 32,203 | 12 | 1 | 1 | ✗ | ✗ | ✓ |
| CityPersons (2017) | 5000 | 6 | 1 | 1 | ✗ | ✗ | ✓ |
| PUCPR+ (2017) | 125 | 135 | 1 | 1 | ✓ | ✓ | ✓ |
| CARPK (2018) | 1448 | 61 | 1 | 1 | ✓ | ✓ | ✓ |
| Open Images V4 (2018) | **1,910,098** | 8.4 | 600 | 2.3 | ✗ | ✓ | ✓ |
| **Our SKU-110K** | 11,762 | **147.4** | **110,712** | **86** | ✓ | ✓ | ✓ |

tinyurl.com/sku110k

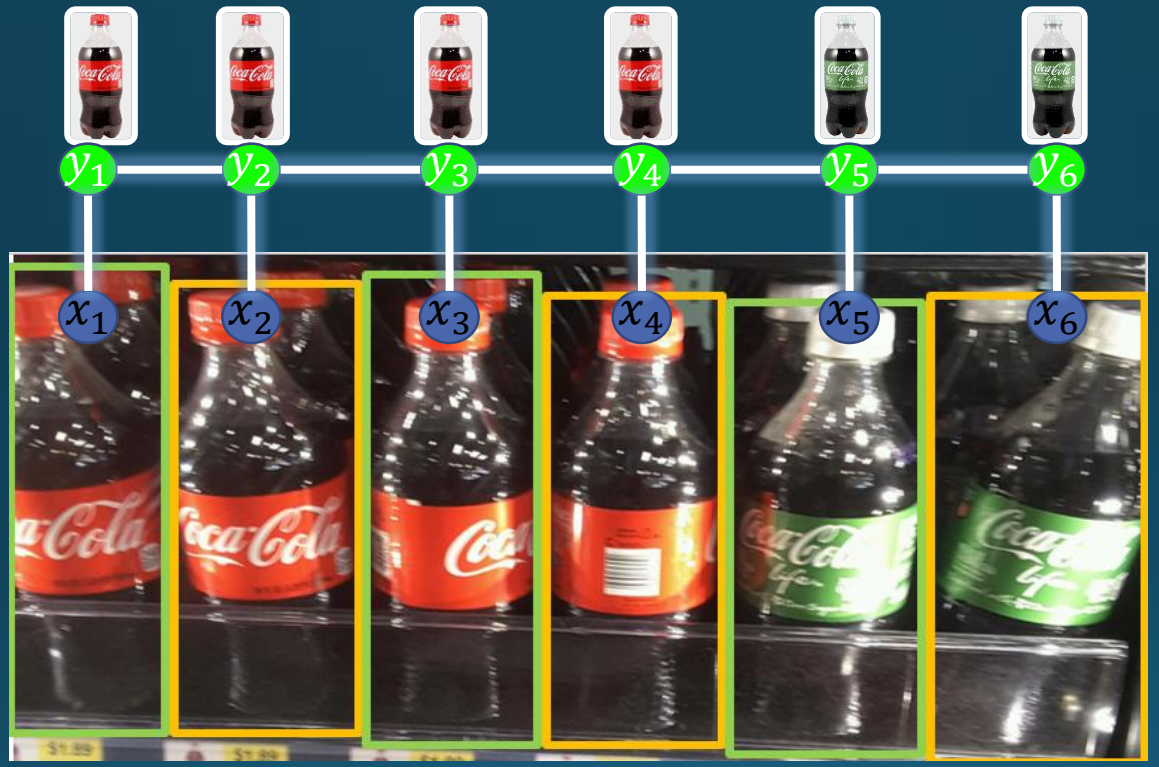

tinyurl.com/sku110k


tinyurl.com/sku110k

# Classification challenges

- *Ultrafine-grained dataset*: classes need context to be recognizable.

# Shelf as a linear-chain CRF



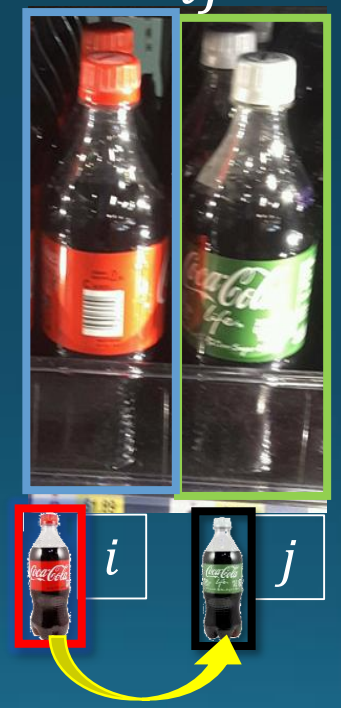

E. Goldman and J. Goldberger, CRF with deep class embedding for large scale classification. Computer Vision and Image Understanding (2019) https://arxiv.org/pdf/1705.07420.pdf

$$P(y_1, y_2, \cdots, y_n | x_1, x_2, \cdots, x_n) = \frac{1}{Z} \prod_{t=2}^{n} \Phi_{i,j}(y_{t-1}, y_t) \prod_{t=1}^{n} \Psi_i(y_t, x_t)$$

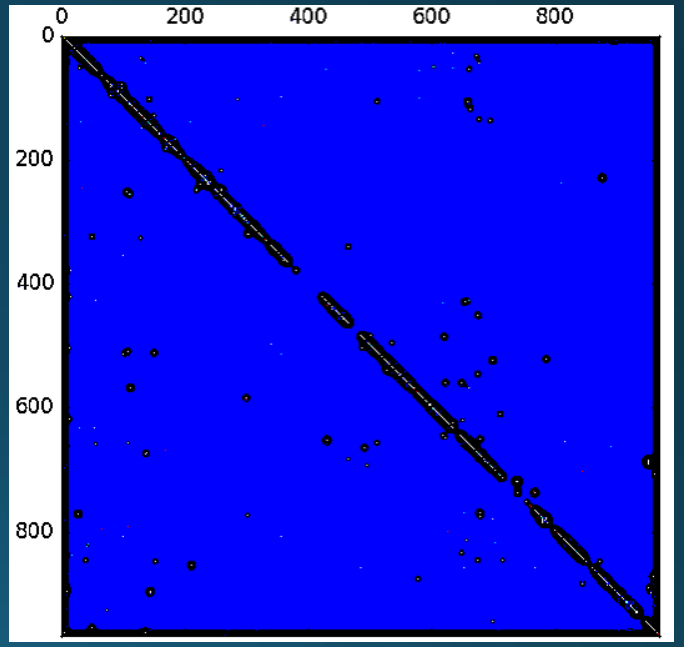
# Log-linear Model

- $P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\prod_{t=2}^{n}\Phi_{i,j}\prod_{t=1}^{n}\Psi_i$
- $\Psi_i(y_t, x_t) = \exp(x_t^T U y_t + y_t^T b)$
- $\Phi_{i,j}(y_{t-1}, y_t) = \exp(y_{t-1}^T P y_t)$
- $P_{ij}$ transition from right to left class
- $P$ is large and sparse

$P_{ij}$




$P_{m\times m}$




millions of parameters

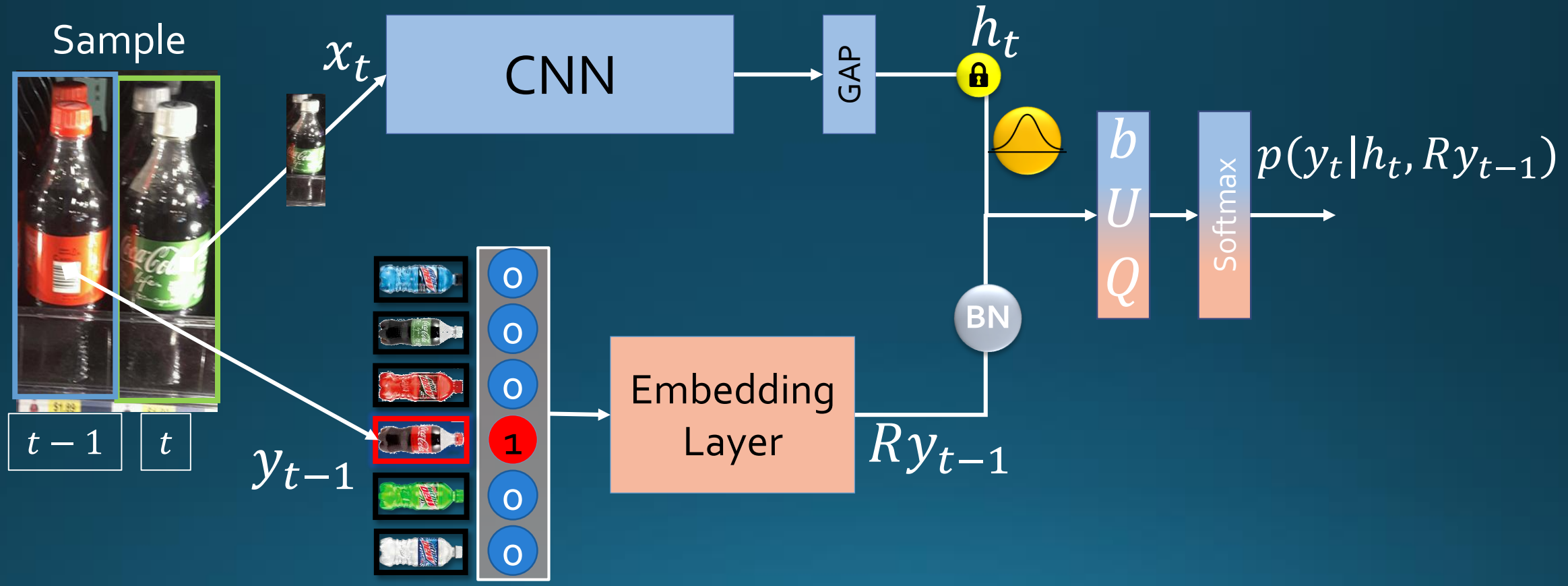
E. Goldman and J. Goldberger, CRF with deep class embedding for large scale classification. Computer Vision and Image Understanding (2019) 102865,https://doi.org/10.1016/j.cviu.2019.102865.

# CRF Pairwise Matrix Decomposition

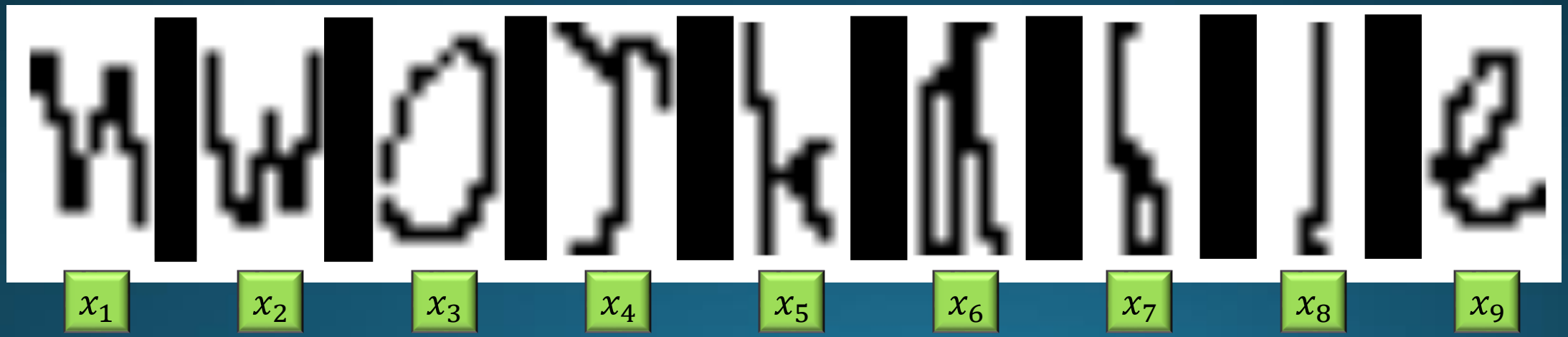
- $P = R^T Q$
- Low-rank matrix factorization
- $2^{nd}$ order semantic similarity of classes
- Non-convex sequence objective
- Hard to optimize


E. Goldman and J. Goldberger, CRF with deep class embedding for large scale classification. Computer Vision and Image Understanding (2019) 102865,https://doi.org/10.1016/j.cviu.2019.102865.

# Approximate MEMM Likelihood




E. Goldman and J. Goldberger, CRF with deep class embedding for large scale classification. Computer Vision and Image Understanding (2019) 102865,https://doi.org/10.1016/j.cviu.2019.102865.


E. Goldman and J. Goldberger, CRF with deep class embedding for large scale classification. Computer Vision and Image Understanding (2019) 102865,https://doi.org/10.1016/j.cviu.2019.102865.

# Our contributions

- ***New detection dataset with new challenges:*** Densely packed retail images
- ***Soft-IoU layer*** estimating overlaps between detections and objects
- ***EM-Merger unit*** resolving overlapping detections by hierarchical clustering
- ***CRF with class embeddings*** classification by both image and context

# Revolutionizing the World of Retail - New Computer Vision Challenges Workshop


More details to follow soon

tinyurl.com/sku110k

# Revolutionizing the World of Retail - New Computer Vision Challenges Workshop


More details to follow soon

tinyurl.com/sku110k