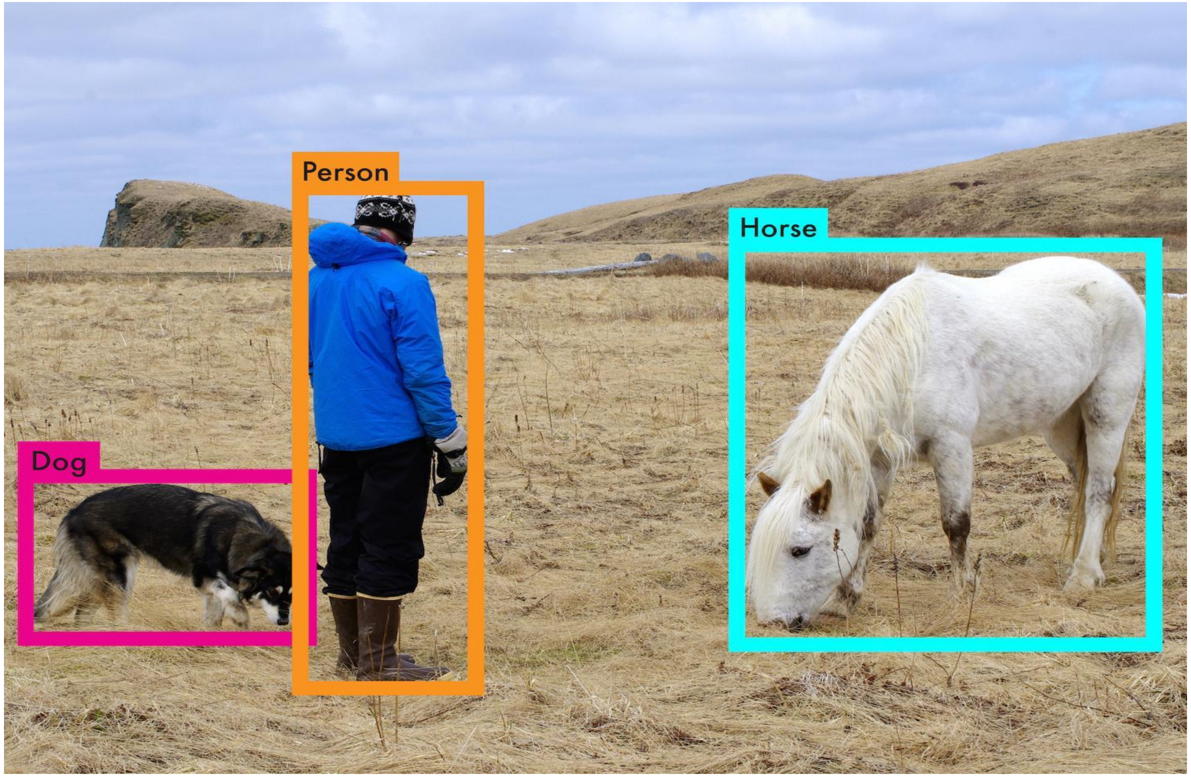
Locating Objects Without Bounding Boxes

Ribera et al. (CVPR 2019)

Utkarsh Ojha, Aron Sarmasi, Rafael A. Rivera-Soto

Object detection



Object detection

Object detection

- Bounding boxes are hard to collect

Object detection

- Bounding boxes are hard to collect
- Why do we even need them?

Object detection

- Bounding boxes are hard to collect
- Why do we even need them?
- Smaller objects seem easier to be detected by **points** than boxes.

Object detection

- Bounding boxes are hard to collect
- Why do we even need them?
- Smaller objects seem easier to be detected by **points** than boxes.
- Points are sometimes enough for weaker localization, and counting instances.

Goal of the paper

Input

Goal of the paper

Input

Predicted locations

Goal of the paper

Input

Predicted locations

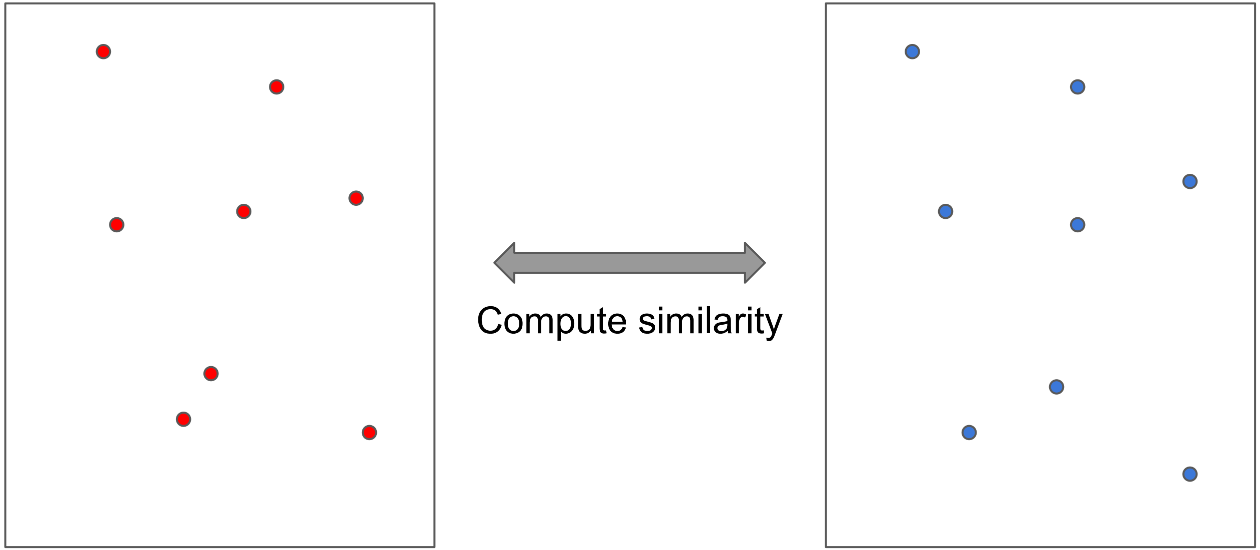
29
(Object count)

Technical contributions

- Modified **Hausdorff loss** to train a *Fully* Convolutional Neural Network (FCN) for object localization.
- Joint estimation of location and number of objects without access to bounding boxes.
- Outperforms state-of-the-art generic object detectors; achieves comparable results for crowd counting.

Localizing points

Object localization ⟷ Estimating point locations

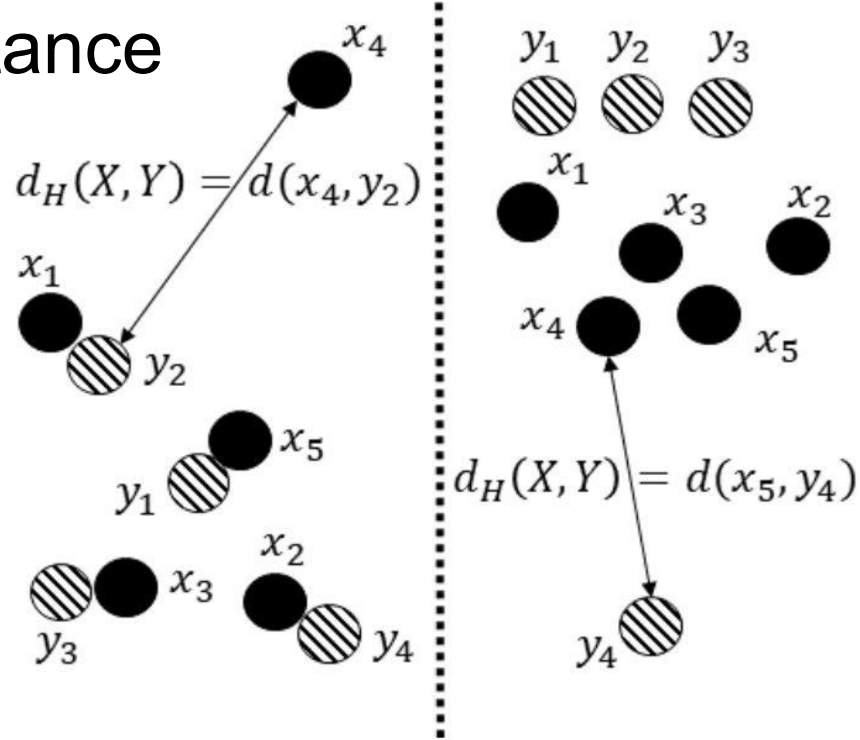


Model predictions

Ground truth

Visualizing Hausdorff Distance

- Largest smallest distance between points in X and Y
- Intuition: measure of distance of worst outlier
- Not a very good measure for point localization
- Not differentiable w.r.t the FCN output



$$d_{\mathrm{H}}(X,Y) = \max\left\{ \sup_{x\in X} \inf_{y\in Y} d(x,y), \sup_{y\in Y} \inf_{x\in X} d(x,y) \right\}$$

Image taken from Ribera et al 2019

Improvements to Hausdorff Distance

- We need a distance measure that is differentiable w.r.t the FCN output p
- Every output pixel/activation needs to contribute to loss
- High activations near ground truth should have little penalty, and low activations far from the closest ground truth should have little penalty

$$d_{\mathrm{WH}}(p, Y) = \underbrace{\frac{1}{\mathcal{S} + \epsilon} \sum_{x \in \Omega} p_x \min_{y \in Y} d(x, y)}_{\text{High loss to activations far from ground truth}} + \underbrace{\frac{1}{|Y|} \sum_{y \in Y} \mathop{M_\alpha}_{x \in \Omega} \left[p_x d(x, y) + (1 - p_x) d_{max} \right]}_{\text{Discourages all-zero activations, as term inside generalized mean is maximized by all-zeros}}$$

Intuition: penalize high activations far from ground truth.

$$d_{\mathrm{WH}}(p, Y) = \frac{1}{\mathcal{S} + \epsilon} \sum_{x \in \Omega} p_x \min_{y \in Y} d(x, y)$$

Intuition: penalize high activations far from ground truth.

Summing over all pixel coordinates

The entire output activation matrix

$$d_{\mathrm{WH}}(p, Y) = \frac{1}{\mathcal{S} + \epsilon} \sum_{x \in \Omega} p_x \min_{y \in Y} d(x, y)$$

Euclidean distance to closest ground truth coordinate

Sum of all activations $\mathcal{S} = \sum_{x \in \Omega} p_x$

Activation at coordinate x

Intuition: discourage all zero activations

$$+ \quad \frac{1}{|Y|} \sum_{y \in Y} \underset{x \in \Omega}{M_\alpha} \left[p_x d(x, y) + (1 - p_x) d_{max} \right]$$

Intuition: discourage all zero activations

Euclidean distance to ground truth point y (not necessarily closest)

Summing over all ground truth points

Maximum possible Euclidean distance (diagonal of the img)

$$+ \quad \frac{1}{|Y|} \sum_{y \in Y} \underbrace{M_{\substack{\alpha \\ x \in \Omega}} \left[p_x d(x,y) + (1 - p_x) d_{max} \right]}_{\text{Generalized mean taken over all pixel coordinates}}$$

Normalize by the number of ground truth points

$$p_x d(x, y) + (1 - p_x) d_{max}$$

Px = 1, close to gt	True positive	low	low	✓
Px = 1, far from gt	False positive	high	low	✓
Px = 0, close to gt	False negative	low	dmax	✓
Px = 0, far from gt	True negative	low	dmax	?

Generalized Mean to the Rescue!

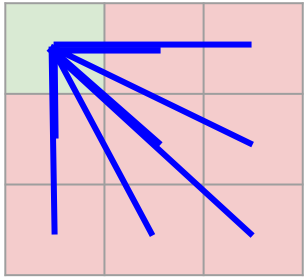
$$\underset{a \in A}{M_\alpha}[f(a)] = \left(\frac{1}{|A|} \sum_{a \in A} f^\alpha(a) \right)^{\frac{1}{\alpha}}$$

where a = $p_x d(x, y) + (1 - p_x) d_{max}$

Num
pixels = n

How can the harmonic mean help?

Ground truth (Y)



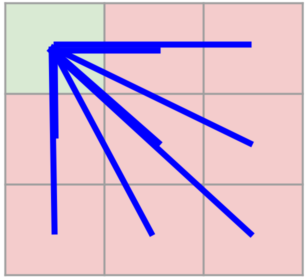
Dmax = 2.83

Predictions (px)

0	1	0
0	0	0
0	0	0

How can the harmonic mean help?

Ground truth (Y)



Predictions (px)

0	1	0
0	0	0
0	0	0

d(x, y)	$p_x d(x, y) + (1 - p_x) d_{max}$
1	1
1	2.83
1.41	2.83
2	2.83
2	2.83
2.23	2.83
2.23	2.83
2.83	2.83

Dmax = 2.83

minimum = 1 (**α** = -∞)

harmonic mean = 1.78 (**α** = -1)

geometric mean = 2.52 (**α** = 0)

arithmetic mean = 2.63 (**α** = 1)

maximum = 2.83 (**α** = +∞)

$$p_x d(x,y) + (1 - p_x) d_{max}$$

Px = 1, close to gt	True positive	low	low	✓
Px = 1, far from gt	False positive	high	low	✓
Px = 0, close to gt	False negative	low	dmax	✓
Px = 0, far from gt	True negative	low	dmax	?

Harmonic mean **greatly weighted** towards the lower values of $p_x d(x,y) + (1 - p_x) d_{max}$

$$\frac{1}{|Y|}\sum_{y\in Y} M_{\alpha \atop x\in\Omega}\left[\,p_x d(x,y) + (1-p_x) d_{max}\,\right]$$

most penalty

least penalty

		Penalty Amount
Px = 1, close to gt	True positive	least
Px = 1, far from gt	False positive	most
Px = 0, close to gt	False negative	most
Px = 0, far from gt	True negative	middle

penalty

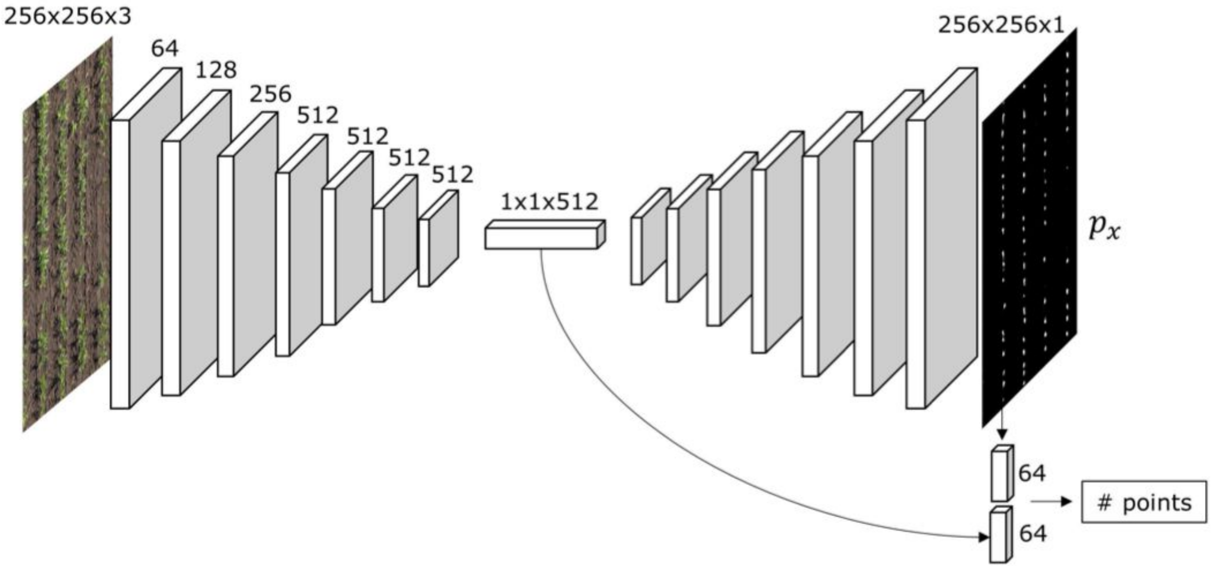
$$\frac{1}{\mathcal{S}+\epsilon}\sum_{x\in\Omega} p_x \min_{y\in Y} d(x,y)$$

To tie it all together:

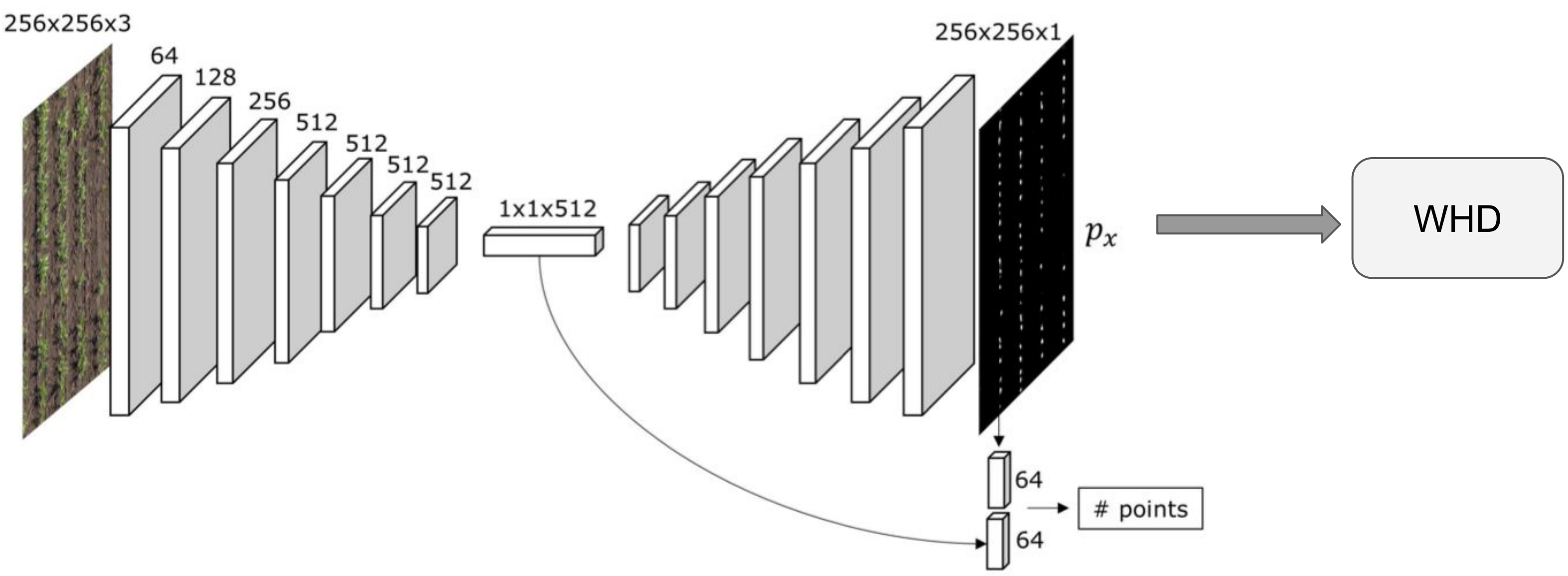
$$d_{\mathrm{WH}}(p,Y) = \frac{1}{\mathcal{S}+\epsilon}\sum_{x\in\Omega} p_x \min_{y\in Y} d(x,y) \; + \; \frac{1}{|Y|}\sum_{y\in Y} \underset{x\in\Omega}{M_\alpha}\left[\, p_x d(x,y) + (1-p_x) d_{max} \right]$$

- Fully differentiable w.r.t output of FCN
- Converges to maximize true positives true negatives, and minimize all else

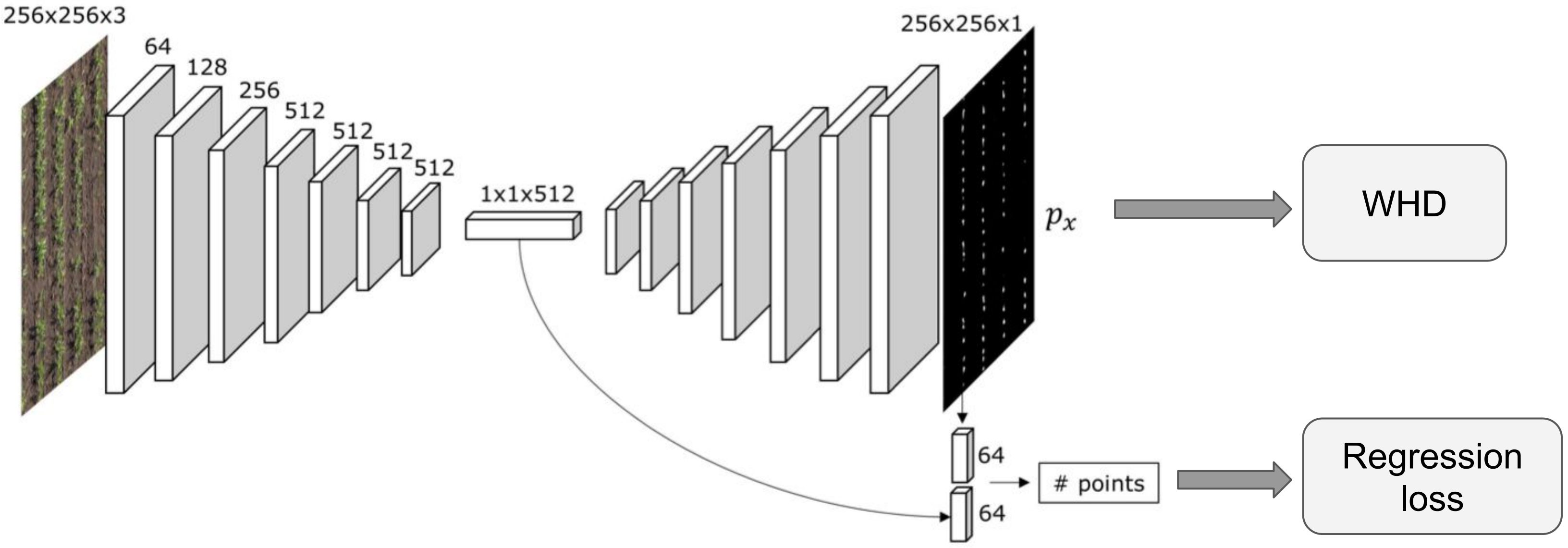
Model Architecture



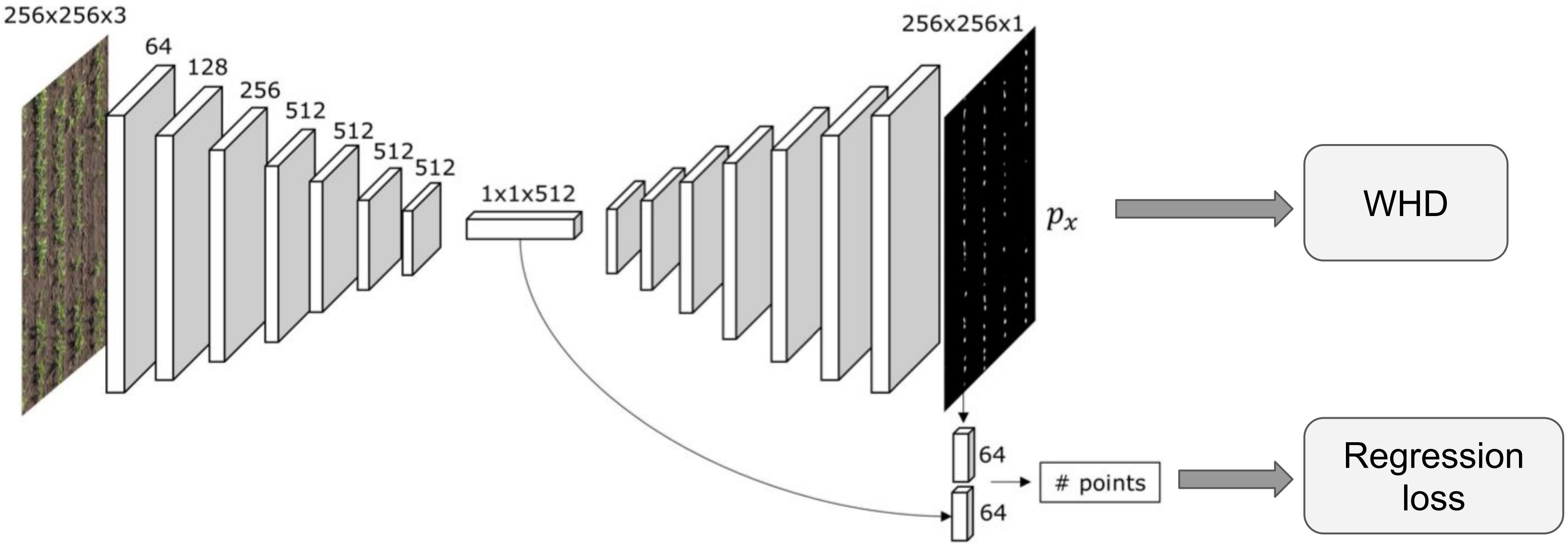
Model Architecture



Model Architecture

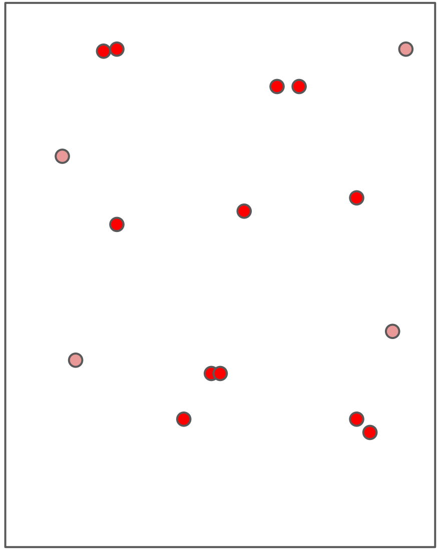


Model Architecture



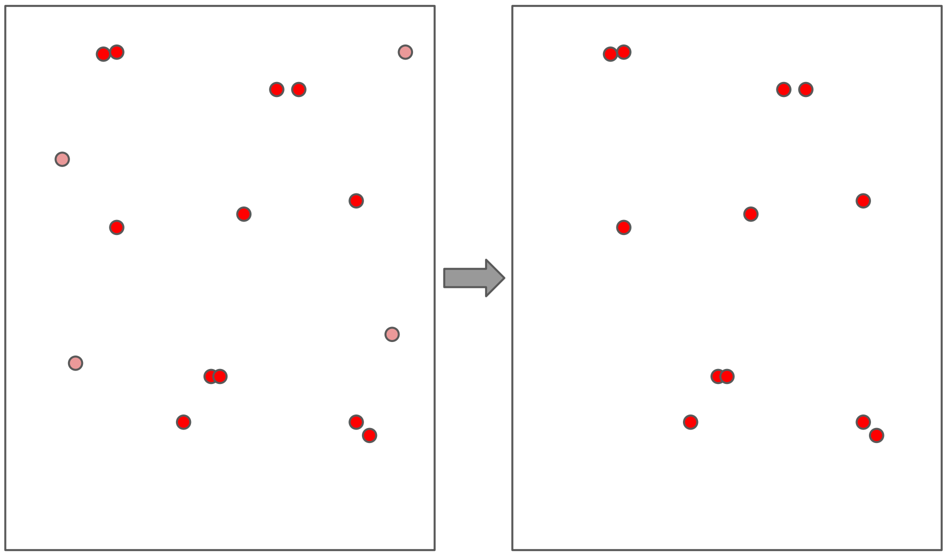
$$L_{reg}(x) = \begin{Bmatrix} 0.5x^2 & \text{for } |x| < 1 \\ |x| - 0.5 & \text{for } |x| \geq 1 \end{Bmatrix}$$

Computing model's predictions



Predicted probability
map

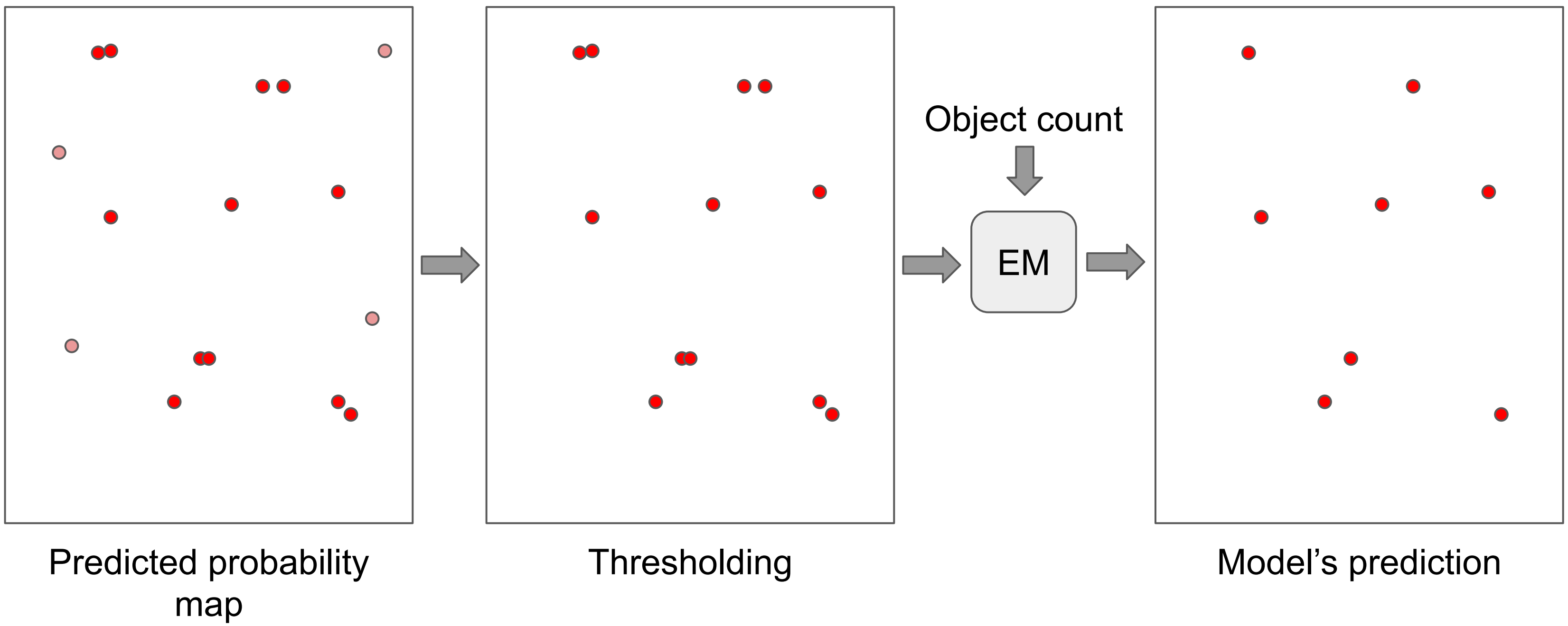
Computing model's predictions



Predicted probability map

Thresholding

Computing model's predictions



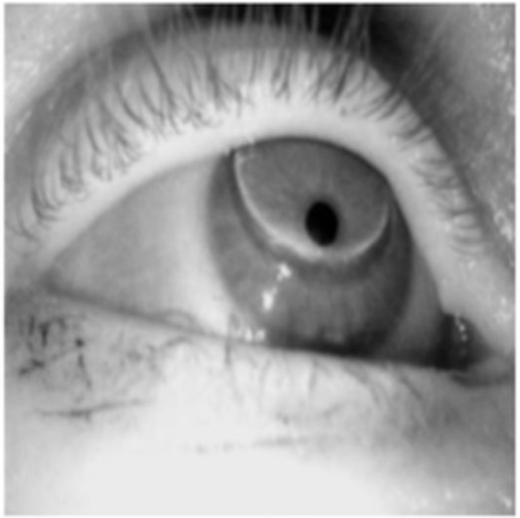
Overall training objective

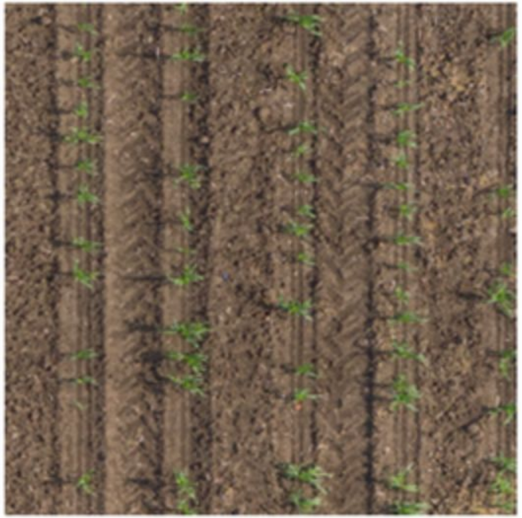
$$L(p, Y) = d_{wh}(p, Y) + L_{reg}(C - \hat{C}(p)))$$

Overall training objective

$$L(p, Y) = d_{wh}(p, Y) + L_{reg}(C - \hat{C}(p)))$$

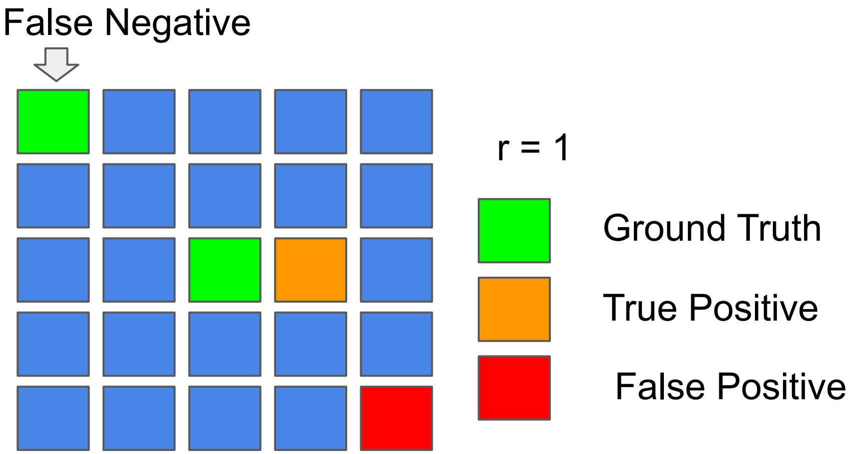
Datasets





- 80/10/10 train, validation and test split for each dataset
- Images resized to 256x256
- Augmented with random horizontal flips

Metrics



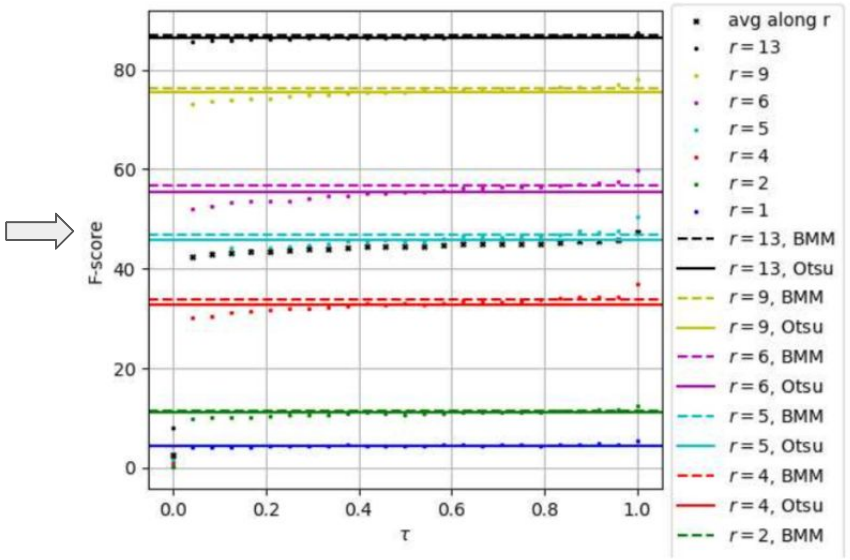
$$Precision = \frac{TP}{TP + FP}$$

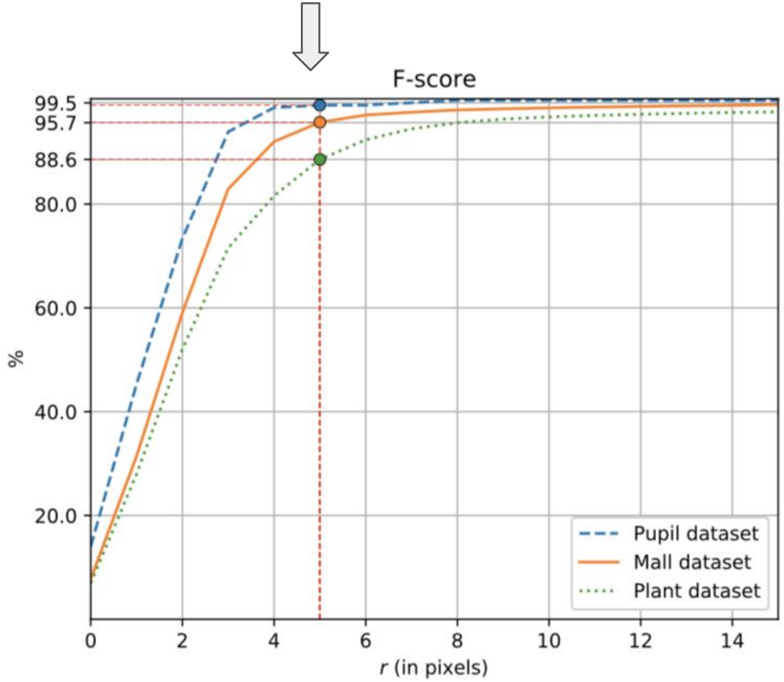
$$Recall = \frac{TP}{TP + FN}$$

$$FScore = 2\frac{Precision * Recall}{Precision + Recall}$$

- Precision and Recall can be 100% even if the model estimates 2 object locations per ground truth point.
- MAE, RMSE and MAPE are reported to counteract this.

Experimental Evaluation

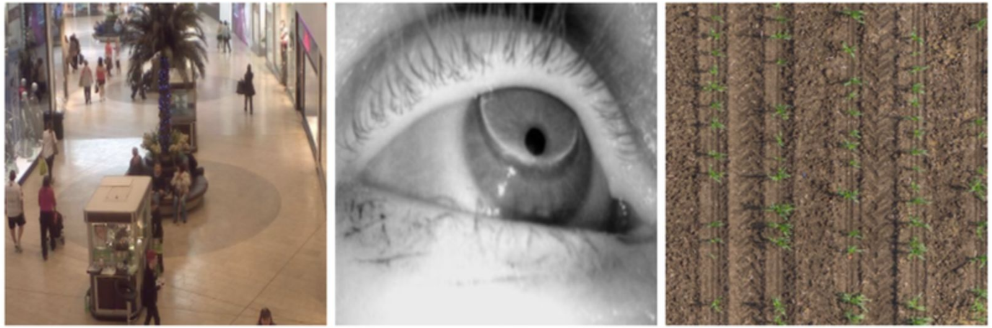




- The bigger “r” is, the easier the problem becomes

Experimental Evaluation

- Comparison against Faster-RCNN with bounding boxes of 20x20 centered at true point
- Model also evaluated on ShanghaiTech Part B achieving MAE of 19.9



Metric	Faster-RCNN	Ours
Precision	81.1%	**95.2 %**
Recall	76.7%	**96.2 %**
F-score	78.8 %	**95.7 %**
AHD	7.6 px	**4.5 px**
MAE	4.7	**1.4**
RMSE	5.6	**1.8**
MAPE	14.8%	**4.4 %**

Method	Precision	Recall	AHD
Swirski [53]	77 %	77 %	-
ExCuSe [13]	77 %	77 %	-
Faster-RCNN	99.5 %	99.5 %	2.7 px
Ours	**99.5** %	**99.5** %	**2.5 px**

Metric	Faster-RCNN	Ours
Precision	86.6 %	**88.1 %**
Recall	78.3 %	**89.2 %**
F-score	82.2 %	**88.6 %**
AHD	9.0 px	**7.1 px**
MAE	9.4	**1.9**
RMSE	13.4	**2.7**
MAPE	17.7 %	**4.2 %**

Strengths

- Dramatically reduces amount of work to annotate a dataset
- No major architectural constraints
- Tested on multiple datasets
- Re-formulation of the object localization problem as the minimization of the distances between a set of points

Weaknesses

- No indication of the size, orientation, occlusion, etc of the object predicted, only center position and instance count
- No comparison between the weighted Hausdorff Distance and other pixel-wise losses such as L2 or MSE.
- Each dataset contained only one type of object, does the method work when trying to detect a wide variety of objects?
- How does it perform with videos, where the objects can exhibit a wide variety of behaviors?
- Notation for generalized mean is misleading.
- Motivation of WHD is done assuming alpha = -inf. But in practice they use alpha = -1.

Thank You!