Human Detection

Purdue University - Vertically Integrated Projects

Image Processing and Analysis - Spring 2023

Robert Sego, Xilai Dai, Alex Weber, Patrick Li, Sun Ahn, Wenjing Chen

Advisors: Prof. Carla Zoltowski, Prof. Edward Delp

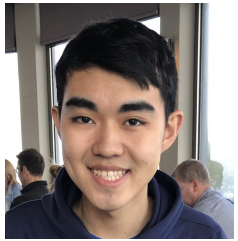
VIP Vertically Integrated Projects

Members

Robert Sego
Computer Engineering

Xilai Dai
Computer Engineering

Alex Weber
Electrical Engineering



Patrick Li
Computer Engineering

Sun Ahn
Computer Science

Human Detection

[Generalizable Pedestrian Detection: The Elephant In The Room]

"we find that existing state-of-the-art pedestrian detectors, though perform quite well when trained and tested on the same dataset, generalize poorly in cross dataset evaluation"

Taken from WIDER Pedestrian Dataset

ETH Pedestrian Dataset

- Contains frames from three kart mounted videos
- Large variety of person sizes
- People are in both background and foreground
- Limited dataset size

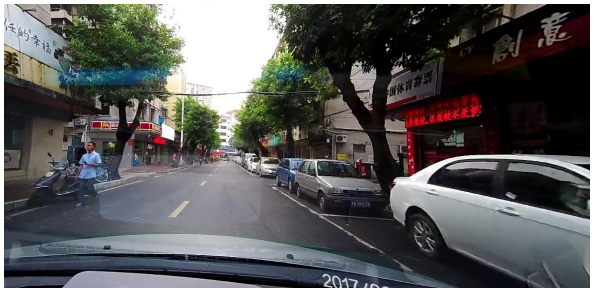
images	1,804
persons	12,298
persons/image	6.81

Table 2: ETH statistics

Link to dataset: [Moving Obstacle Detection in Highly Dynamic Scenes (ethz.ch)]

WIDER 2019 Dataset

- Two Sources: surveillance camera and dashcam
- Both are of urban environments
- Multiple Weather conditions
- Includes both training and validation datasets



images	91,500
persons	292,890
persons/image	3.2

Table 2: WIDER 2019 Training statistics

images	5,000
persons	20,052
persons/image	4.0

Table 3: WIDER 2019 Validation statistics

Link to dataset: CodaLab - Competition

Eurocities Dataset

- Images from multiple sources, (mainly dash camera)
- Contains all seasons, time, and weather conditions
- Large diversity in crowd size
- Annotations of pedestrians, cyclists, motorcycle riders and people

images	47,300
persons	238,200
persons/image	5.03

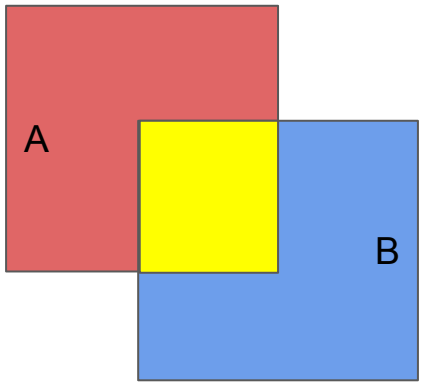
Table 1: ECP statistics

Link to dataset:
https://eurocity-dataset.tudelft.nl/

IOU

$$IOU = \frac{|A \cap B|}{|A \cup B|}$$

A : ground truth box
B : model-output box



Confusion Matrix

		Actual	
		Positive	Negative
Model output	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

True Positives (TP): The model predicted a label and matches correctly as per ground truth.

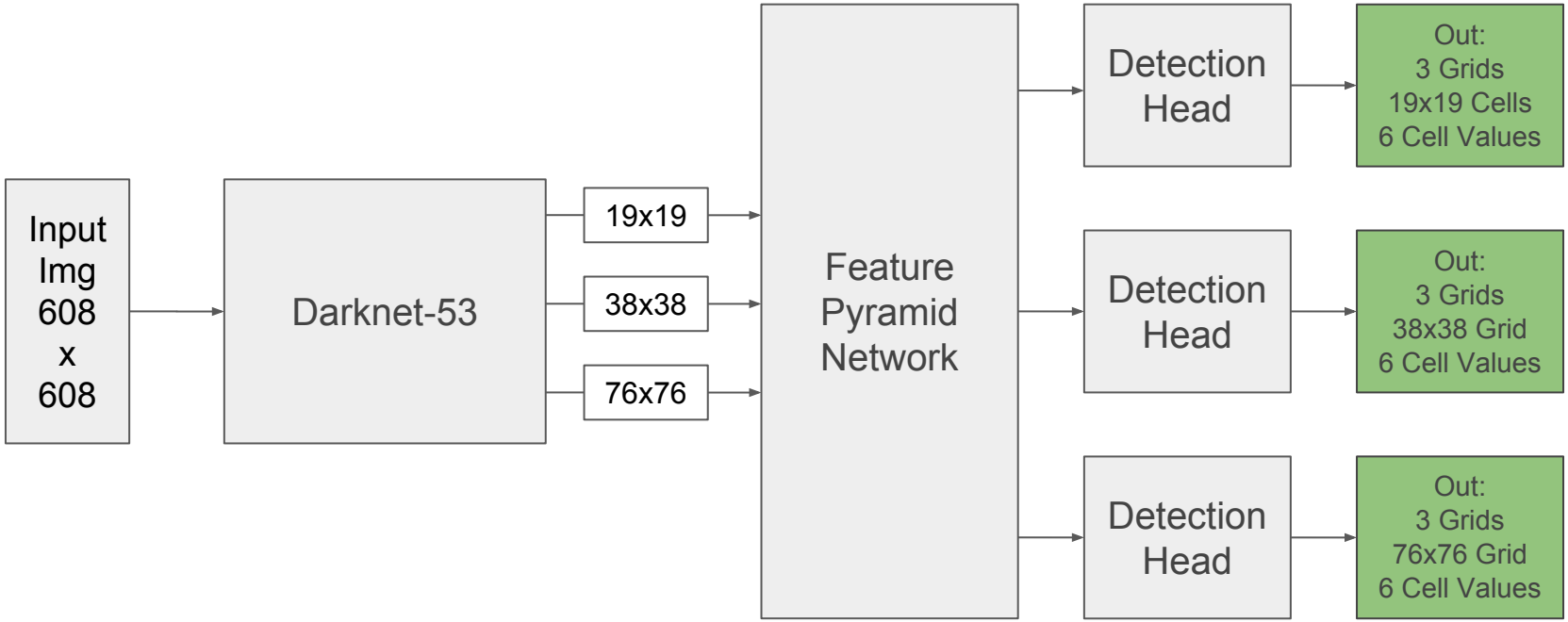
True Negatives (TN): The model does not predict the label and is not a part of the ground truth.

False Positives (FP): The model predicted a label, but it is not a part of the ground truth.

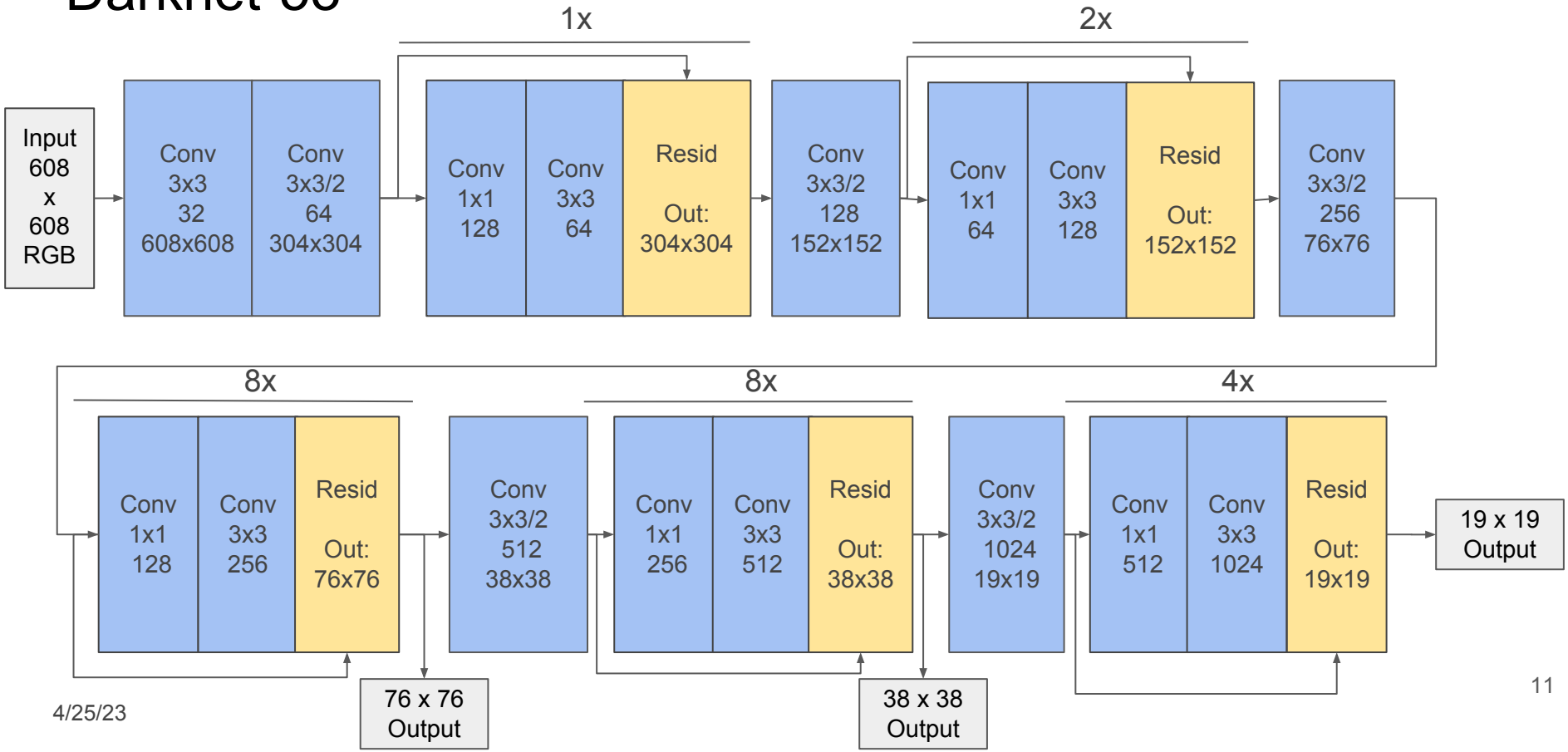
False Negatives (FN): The model does not predict a label, but it is part of the ground truth.

Method 1: YOLO v3

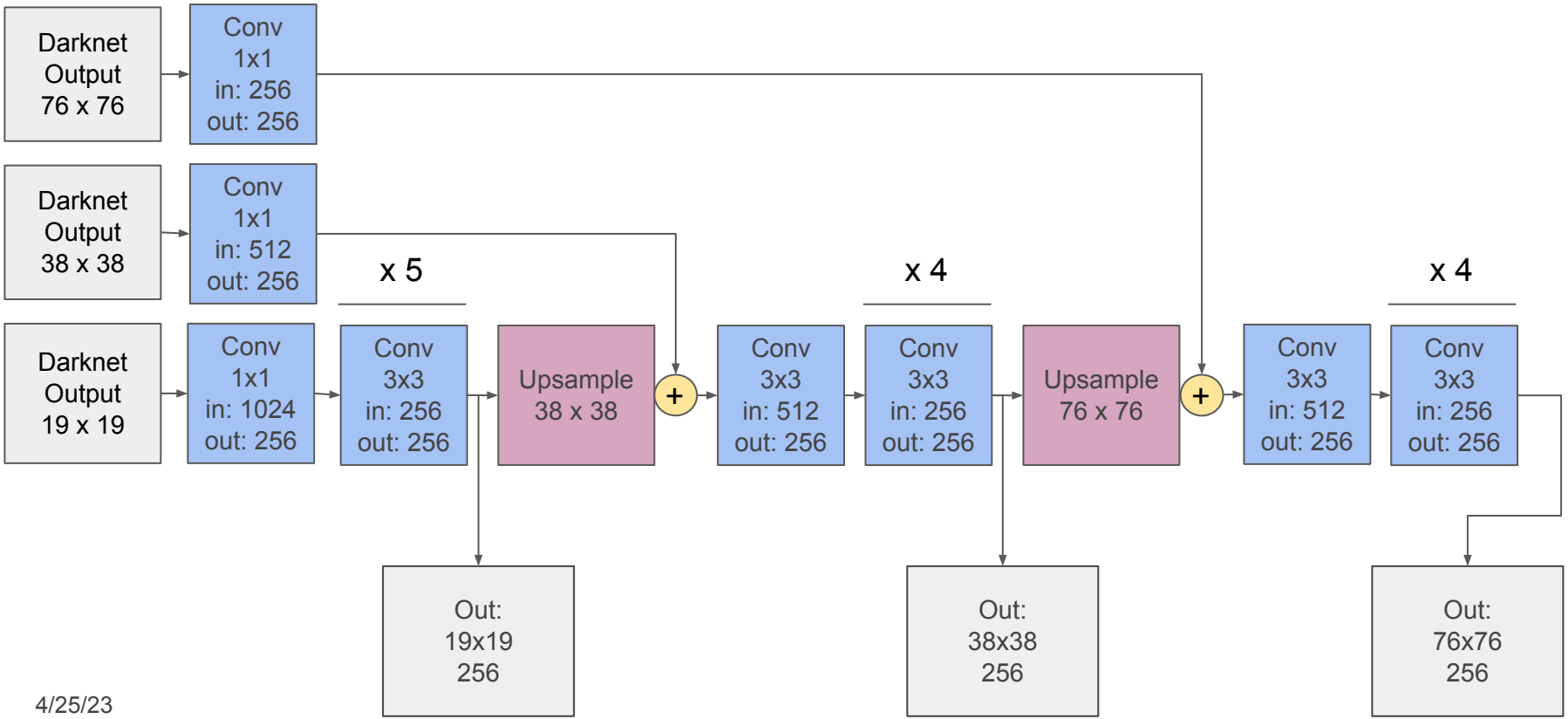
YOLOv3 Structure



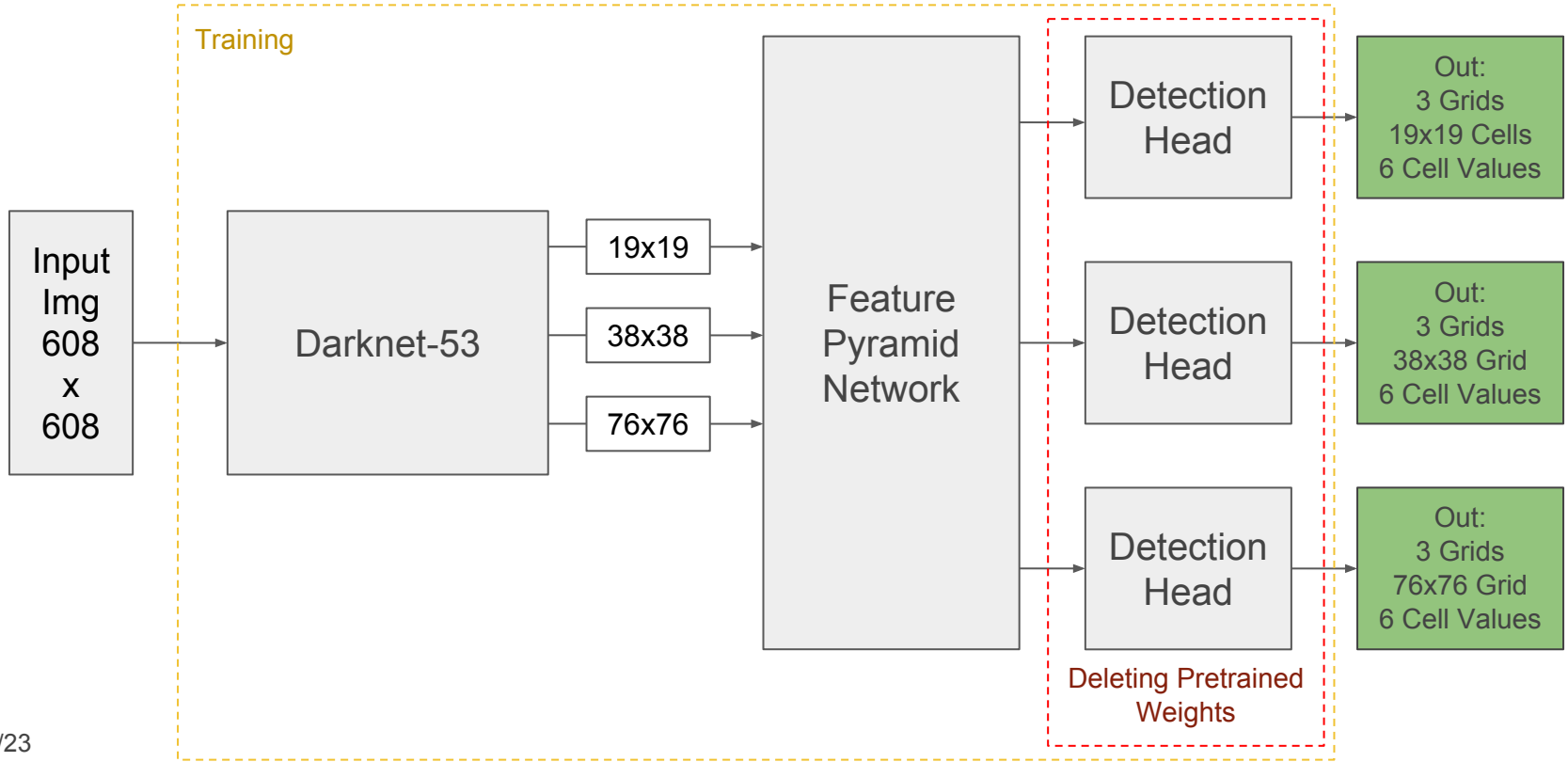
Darknet-53


1x
2x
Input
608
x
608
RGB
Conv
3x3
32
608x608
Conv
3x3/2
64
304x304
Conv
1x1
128
Conv
3x3
64
Resid
Out:
304x304
Conv
3x3/2
128
152x152
Conv
1x1
64
Conv
3x3
128
Resid
Out:
152x152
Conv
3x3/2
256
76x76
8x
8x
4x
Conv
1x1
128
Conv
3x3
256
Resid
Out:
76x76
Conv
3x3/2
512
38x38
Conv
1x1
256
Conv
3x3
512
Resid
Out:
38x38
Conv
3x3/2
1024
19x19
Conv
1x1
512
Conv
3x3
1024
Resid
Out:
19x19
19 x 19
Output
76 x 76
Output
38 x 38
Output

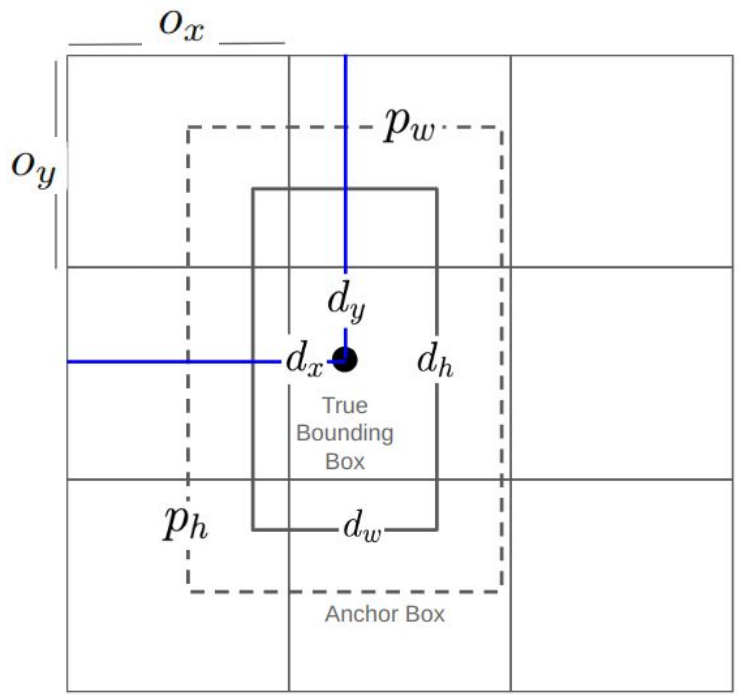
Feature Pyramid Network



YOLO v3 Training Structure



Anchor Boxes



$$d_x = \sigma(t_x) + o_x$$
$$d_y = \sigma(t_y) + o_y$$
$$d_w = p_w e^{t_w}$$
$$d_h = p_h e^{t_h}$$

o_x: gridcell horizontal offset
o_y: gridcell vertical offset
p_w: anchor box width
p_h: anchor box height
d_x, d_y, d_w, d_h: true bounding box coordinates and dimensions
t_x, t_y, t_w, t_h: model outputs

Anchor Boxes - Calculation

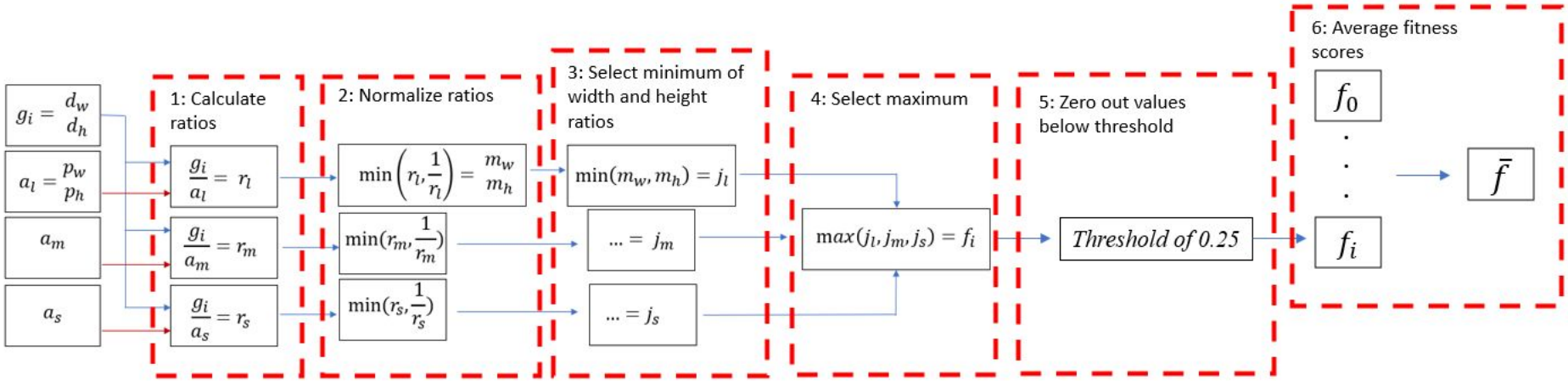
$$a_l = \begin{matrix} p_w \\ p_h \end{matrix} = \begin{matrix} b_w + \sigma_w \\ b_h + \sigma_h \end{matrix}$$

$$a_m = \begin{matrix} b_w \\ b_h \end{matrix}$$

$$a_s = \begin{matrix} b_w - \sigma_w \\ b_h - \sigma_h \end{matrix}$$

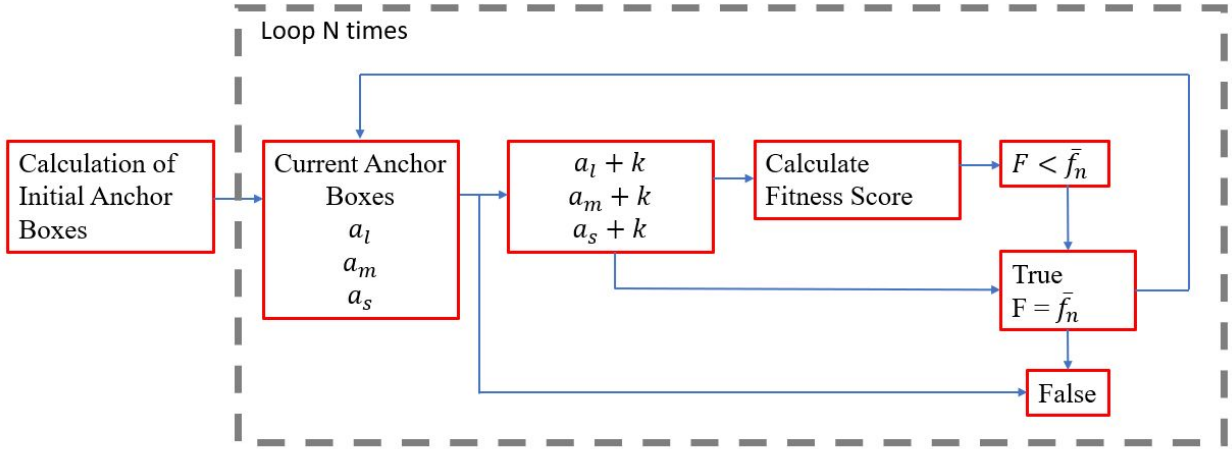
a_l: Large anchor box
a_m: Medium anchor box
a_s: Small anchor box
p_w: Width of anchor box
p_h: Height of anchor box
b_w: Average width of ground truth bounding boxes
b_w: Average height of ground truth bounding boxes
σ_w: Standard deviation of the width of ground truth bounding boxes
σ_h: Standard deviation of the height of ground truth bounding boxes

Anchor Boxes - Fitness



j_l: minimum of normalized width and height ratios from step 2 for large anchor boxes
j_m: minimum of normalized width and height ratios from step 2 for medium anchor boxes
j_s: minimum of normalized width and height ratios from step 2 for small anchor boxes
m_w: normalized width ratio
m_h: normalized height ratio
d_w: bounding box width
d_h: bounding box height
f_i: fitness score for the ith bounding box
$\bar{f}$: average fitness score

Anchor Boxes - Training



a_l: Large anchor box
a_m: Medium anchor box
a_s: Small anchor box
k: Random integer between -10 and 10
F: Current highest fitness score
$\bar{f}_n$: Average fitness score of nth iteration

YOLOv3 Loss

$$\sum_{i=1}^{S^2}\sum_{j=1}^{B} C_{IOU} + \sum_{i=1}^{S^2}\sum_{j=1}^{B} I_{ij}^{obj} FL(C_i) + \sum_{i=1}^{S^2}\sum_{j=1}^{B} I_{ij}^{noobj} FL(C_i) + \sum_{i=1}^{S^2}\sum_{j=1}^{B} I_{ij}^{obj} FL(p_{ij}(c))$$

C_{IOU}: Complete IOU Loss
FL: Focal Loss
I^{obj}: 1 if ground truth contains bounding box in gridcell, 0 otherwise
I^{noobj} : 0 if ground truth contains bounding box in gridcell, 1 otherwise
$p(c)$: confidence that object is of specific class
C: confidence score of presence of an object, 0 or 1 for ground truth

CIOU Loss

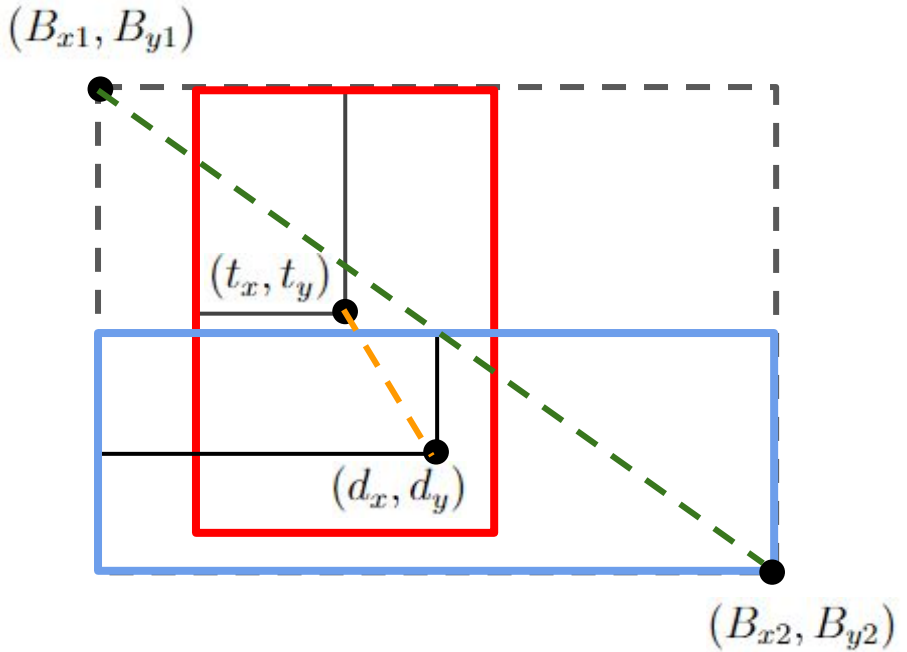
$$C_{IOU} = (1 - IOU) + D_{IOU} + R_{IOU}$$

C_{IOU}: Complete IOU Loss
IOU: Intersection over Union
D_{IOU}: Distance Loss
R_{IOU}: Aspect Ratio Consistency

Distance Loss

$$B_{x1} = argmin(t_x - 0.5t_w, d_x - 0.5d_w)$$
$$B_{x2} = argmax(t_x + 0.5t_w, d_x + 0.5d_w)$$
$$B_{y1} = argmin(t_y - 0.5t_h, d_y - 0.5d_h)$$
$$B_{y2} = argmax(t_y + 0.5t_h, d_y + 0.5d_h)$$

$$D_{IOU} = \frac{\sqrt{(t_x - d_x)^2 + (t_y - d_y)^2}}{\sqrt{(B_{x1} - B_{x2})^2 + (B_{y1} - B_{y2})^2}}$$

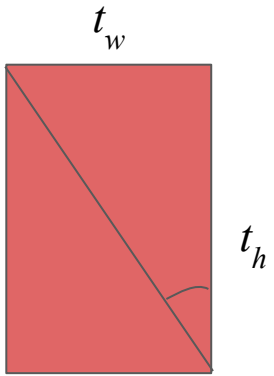


d_x, d_y: true bounding box center
t_x, t_y: model outputs
D_{IOU}: Distance Loss

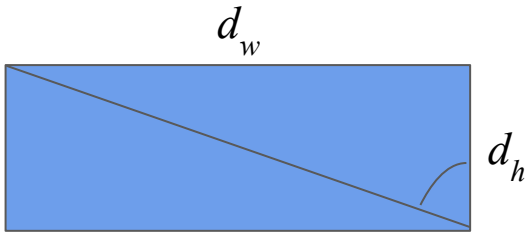
Aspect Ratio Consistency

$$R_{IOU} = \frac{4}{\pi}(\arctan\frac{t_w}{t_h} - \arctan\frac{d_w}{d_h})^2$$

d_w, d_h: true bounding box and dimensions
t_w, t_h: model outputs
R_{IOU}: Aspect Ratio Consistency



Output

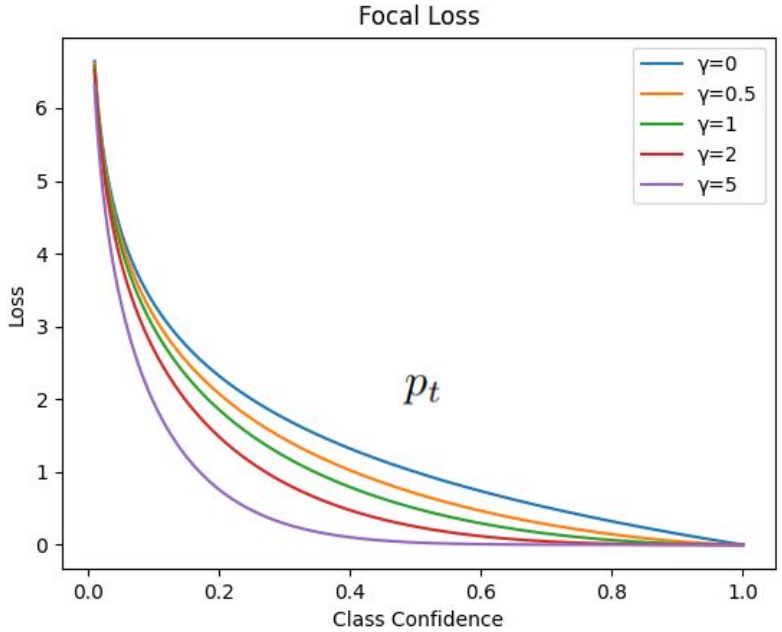


Ground Truth

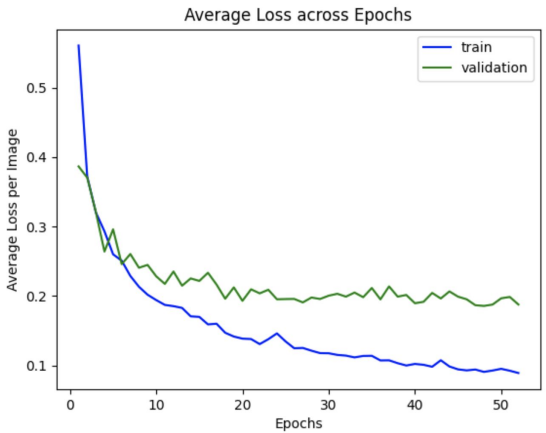
Focal Loss

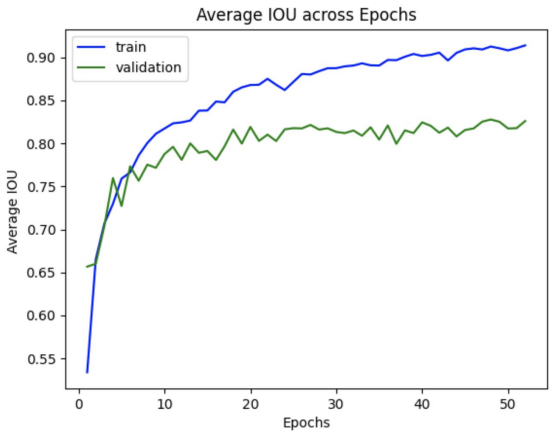
$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log_2(p_t)$$

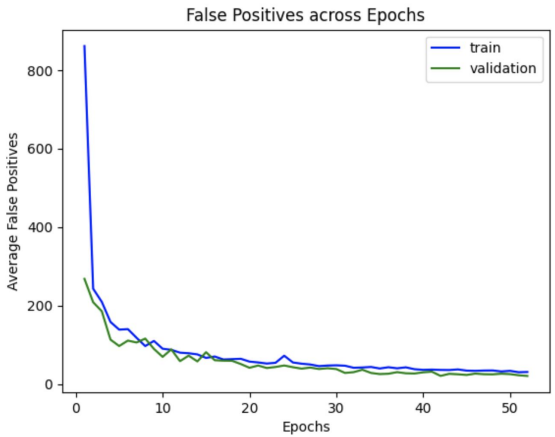
p: model output probability
p_t: p if ground truth is 1, (1-p) if ground truth is 0
γ: focusing parameter
α_t: weighting factor parameter



Training/Validation on ETH

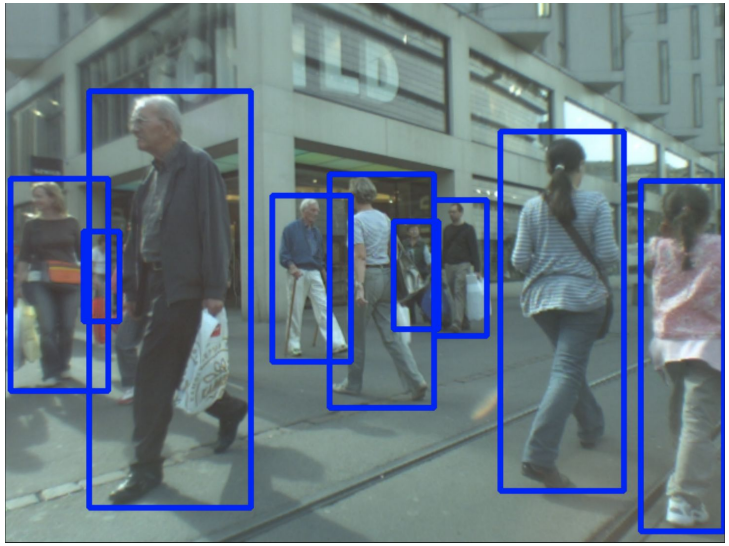




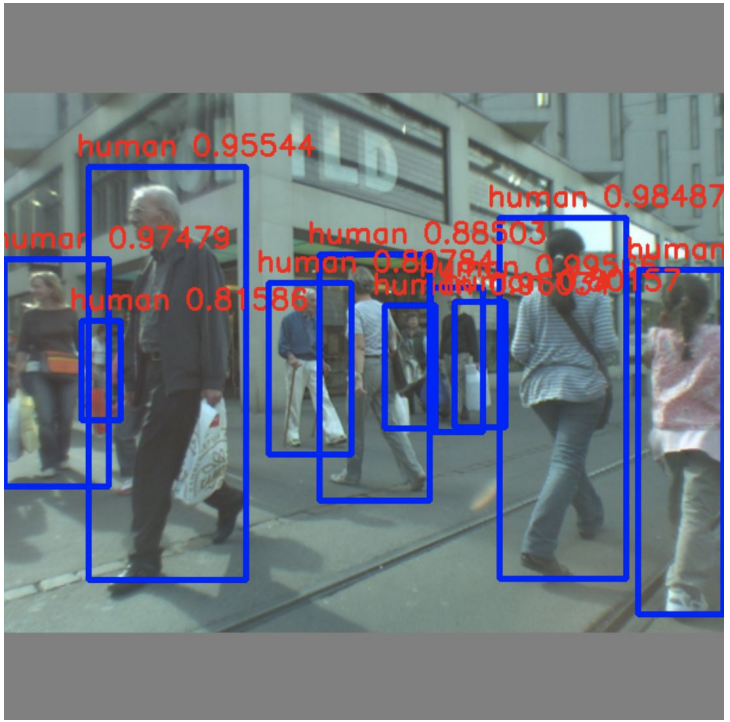


Results for Training Dataset

Ground Truth

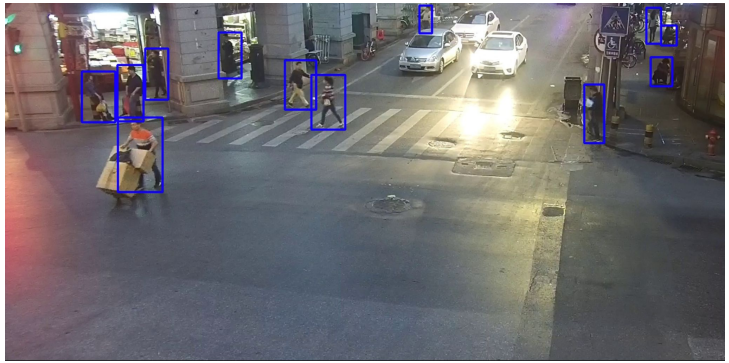


Output

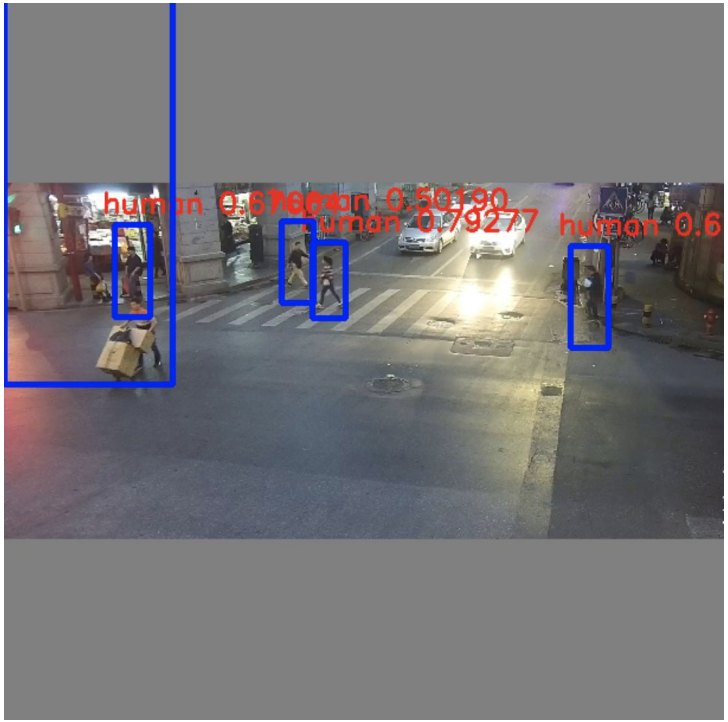


Results for Testing Dataset

Ground Truth

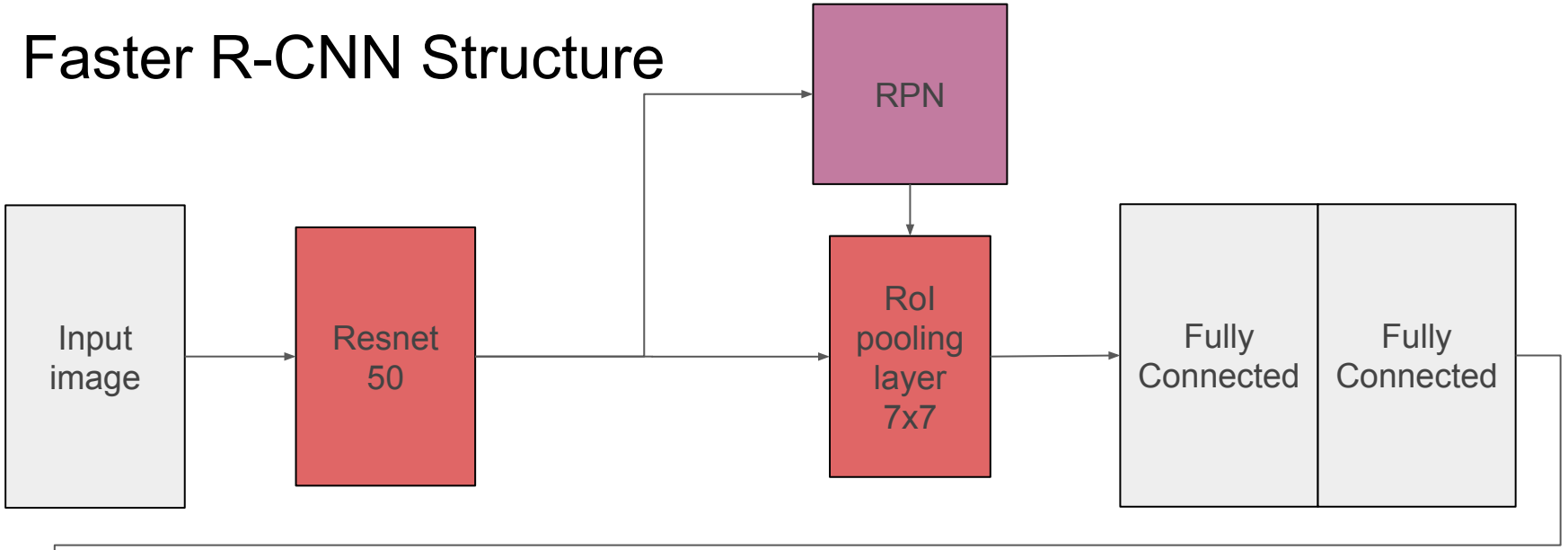


Output



Method 2: Faster R-CNN

Faster R-CNN Structure



$$p = (p_1, p_2, \ldots p_k)$$

p = *predicted class probability*

k: *number of classes*

(t_x^k, t_y^k): *top left bounding box coordinate for k class object*

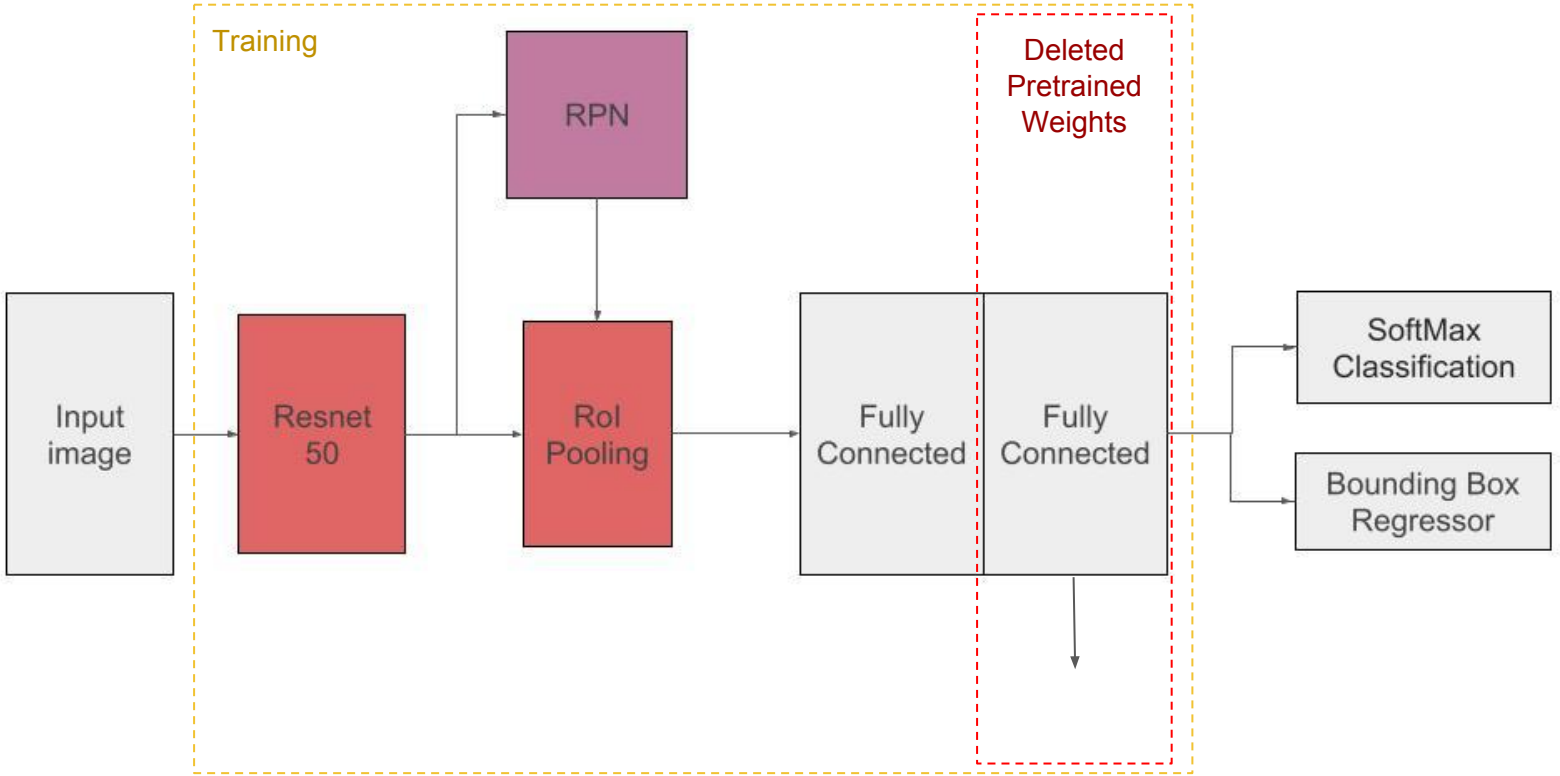
(t_w^k, t_h^k): *bounding box width and height*

$$t^k = (t_x^k, t_y^k, t_w^k, t_h^k)$$

SoftMax classification layer: choose which object class found

Bounding Box Regressor: Output bounding box coordinates for each object class

Pretrained Weights and Training Chart



Loss

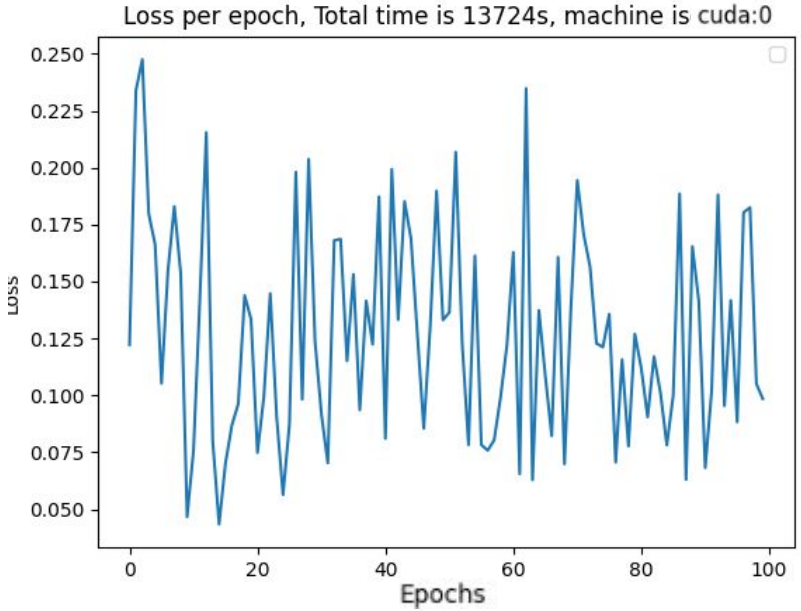
RPN Loss *Faster RCNN Loss*

$$L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) + L_{cls}(p_i, {p_i}^*) + \lambda L_{reg}(t_i, {t_i}^*)$$

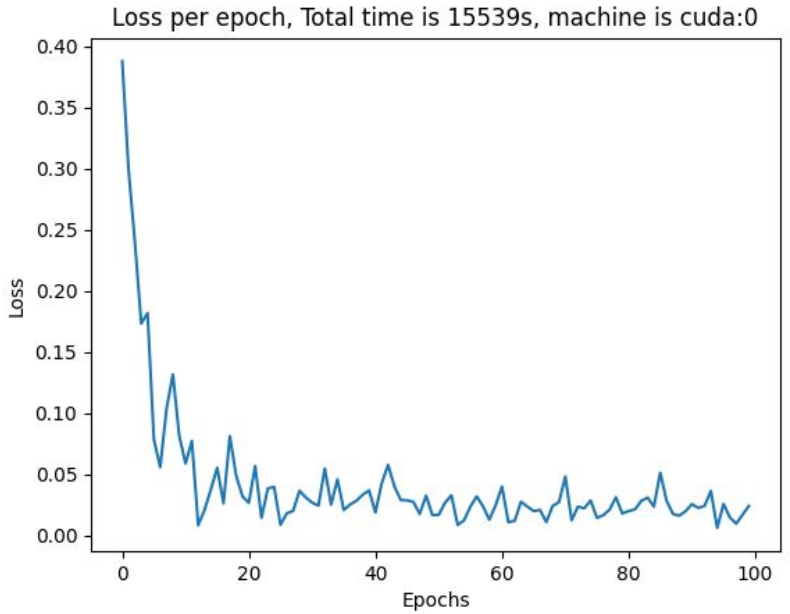
N_{cls}: batch size
N_{reg}: number of anchor locations
λ: parameter coefficient (default=10)
L_{cls}: classification loss
L_{reg}: bounding box regression loss

i: anchor index
p_i: output anchor object confidence
p_i^*: ground truth object label
t_i: output bounding box values

Training/Validation



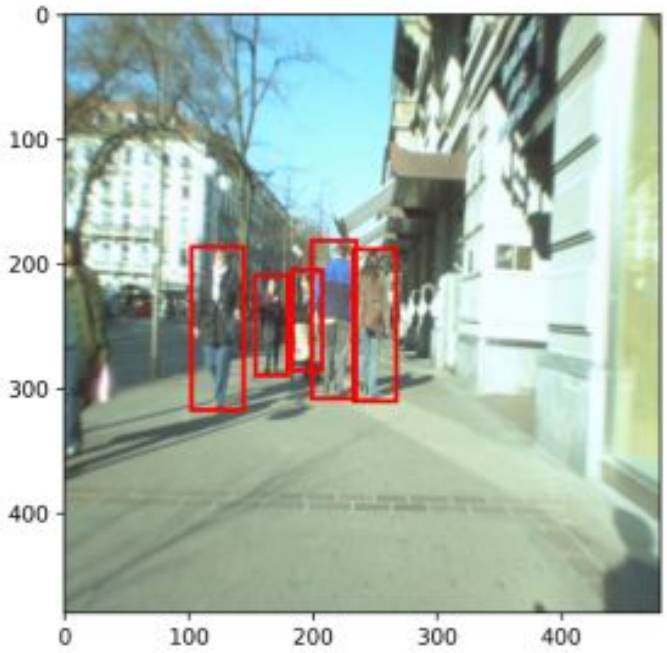
Loss with learnable learning rate



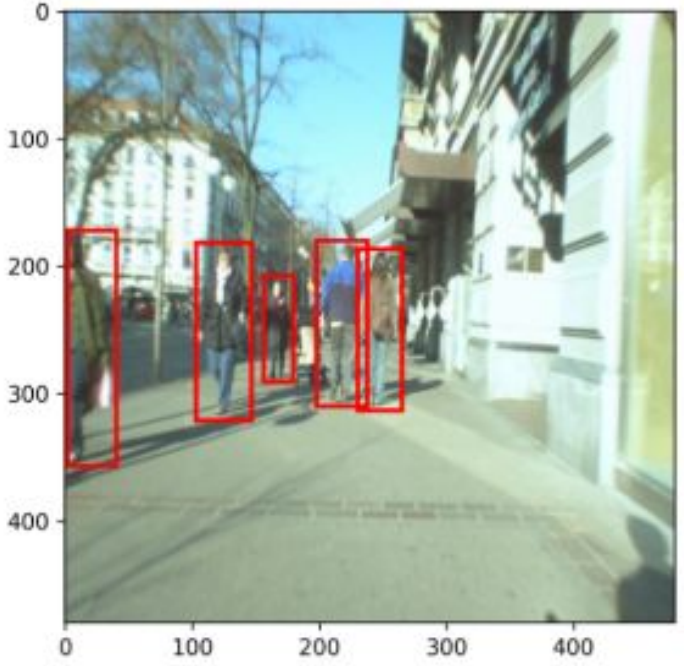
Loss with fixed learning rate = 0.005

Results for Training Dataset

IoU threshold is 0.5



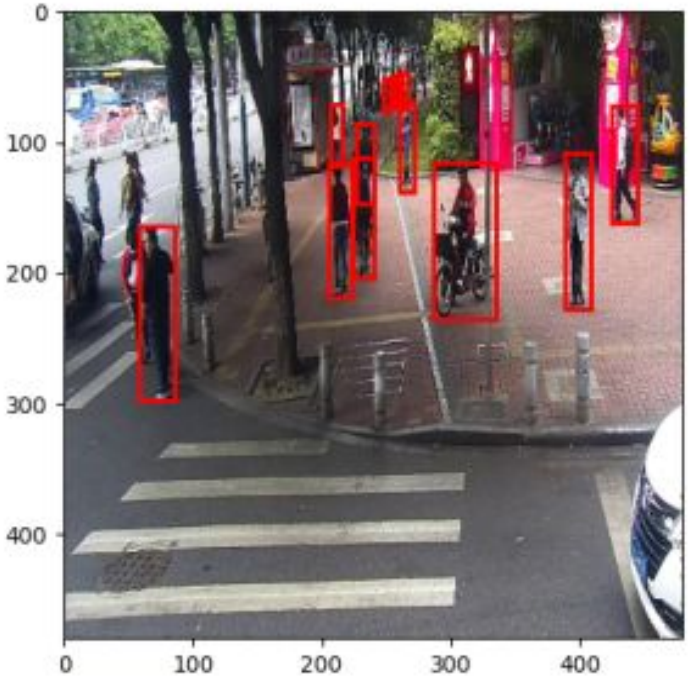
Ground truth of the ETH sample



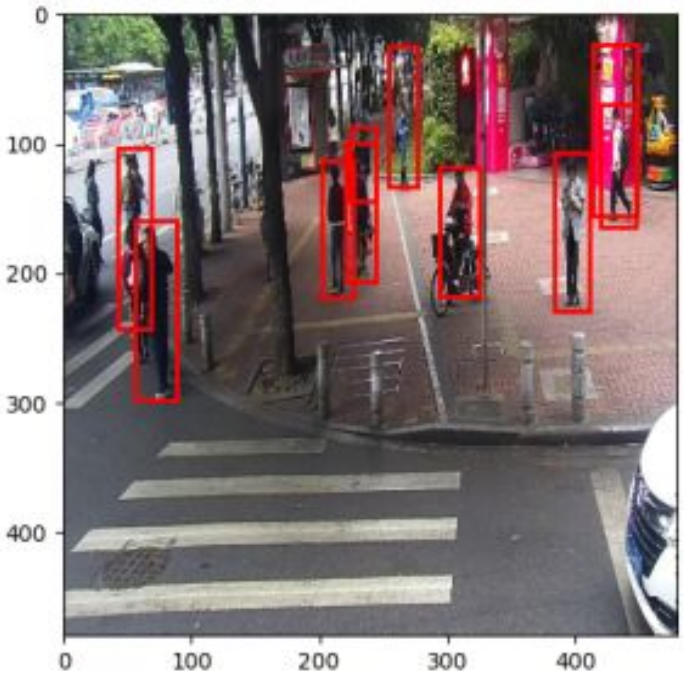
Model output of the ETH sample

Results for Testing Dataset

IoU threshold is 0.5



Ground truth of the Wider sample



Model output of the Wider sample

Conclusion and Comparison

The WIDER 2019 dataset is a varied dataset that acts as a thorough test of a model's generalizability. The ETH dataset does provide a large variance in distance from the camera, but models trained on it do not easily detect people far in the background.

Pretrained YOLO v3 learns human detection datasets relatively quickly, but may more easily overfit than contemporary models. Faster R-CNN requires more training epochs, but generally performs better in cross-dataset evaluation.

Future Work

- Compare relevant pedestrian detectors like Pedestron and F2Dnet
- Implement later versions of YOLO and other object detectors for comparison
- Use Mean Average Precision to evaluate the effectiveness of each model
- Test on other datasets

Sources

[1] Hasan, I., Liao, S., Li, J., Akram, S. U., & Shao, L. (2020, December 9). Generalizable Pedestrian Detection: The Elephant In The Room. *arXiv*. Retrieved November 13, 2022, from https://arxiv.org/pdf/2003.08799.pdf

[2] WIDER Face & Person Challenge 2019 - Track 2: Pedestrian Detection (n.d.). *CodaLab*. Retrieved November 13, 2022, from https://competitions.codalab.org/competitions/20132#learn_the_details-overview

[3] A. Ess and B. Leibe and K. Schindler and and L. van Gool (2008, June). A Mobile Vision System for Robust Multi-Person Tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. *IEEE Xplore*. Retrieved November 13, 2022, from https://data.vision.ee.ethz.ch/cvl/aess/dataset/

[4] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

Sources

[5] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*. https://doi.org/10.48550/arXiv.1506.01497

[6] Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision* (1440-1448). https://doi.org/10.48550/arXiv.1504.08083

[7] Redmon, J., & Farhadi, A. (2018, April 8). *Yolov3: An incremental improvement*. arXiv.org. https://doi.org/10.48550/arXiv.1804.02767