An active learning framework to optimize training of deep models with human-in-the-loop

Training with less labels

Humayun Irshad, Lead Scientist
Figure Eight

HOW **BIG** IS BIG DATA?

We need to find better and more efficient ways to label and use our data

Supervised Learning

Data

Traditional → BIG

DATA COLLECTION

- Manual labeling
- Random sampling

DATA COLLECTION

- Selective labeling
- Active learning

Predictive Modeling

Traditional → Deep learning

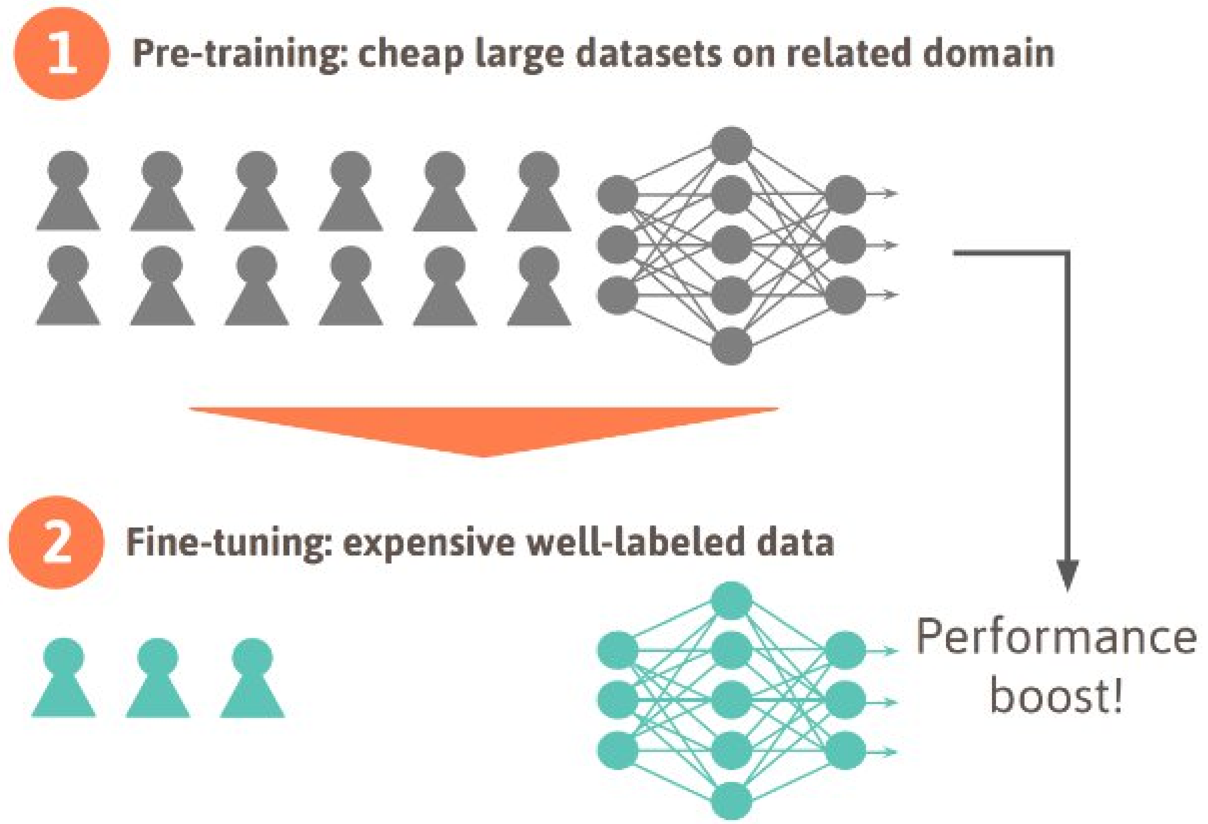
FEATURE EXTRACTION AND CLASSIFICATION

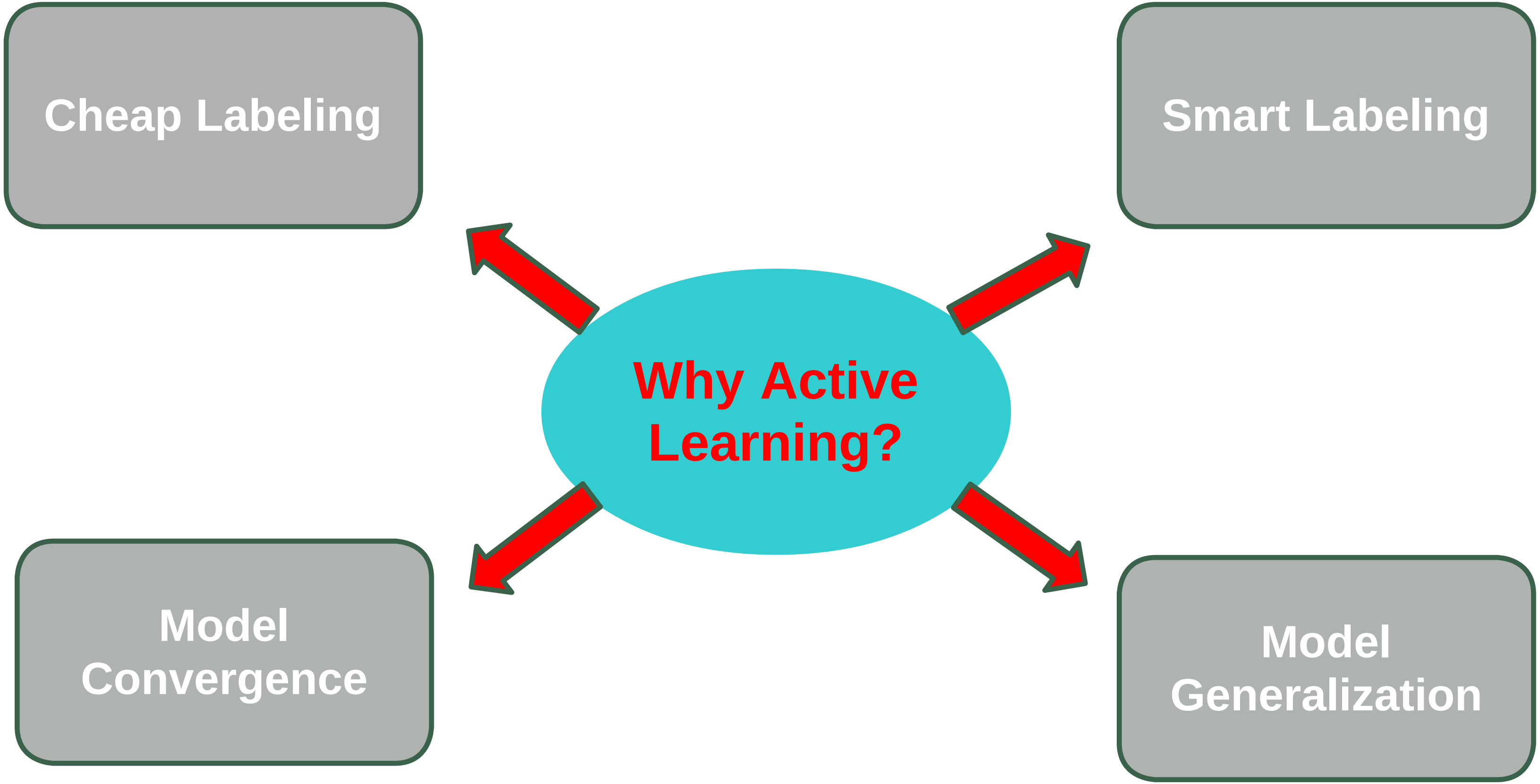
- SVM
- Random Forest
- …

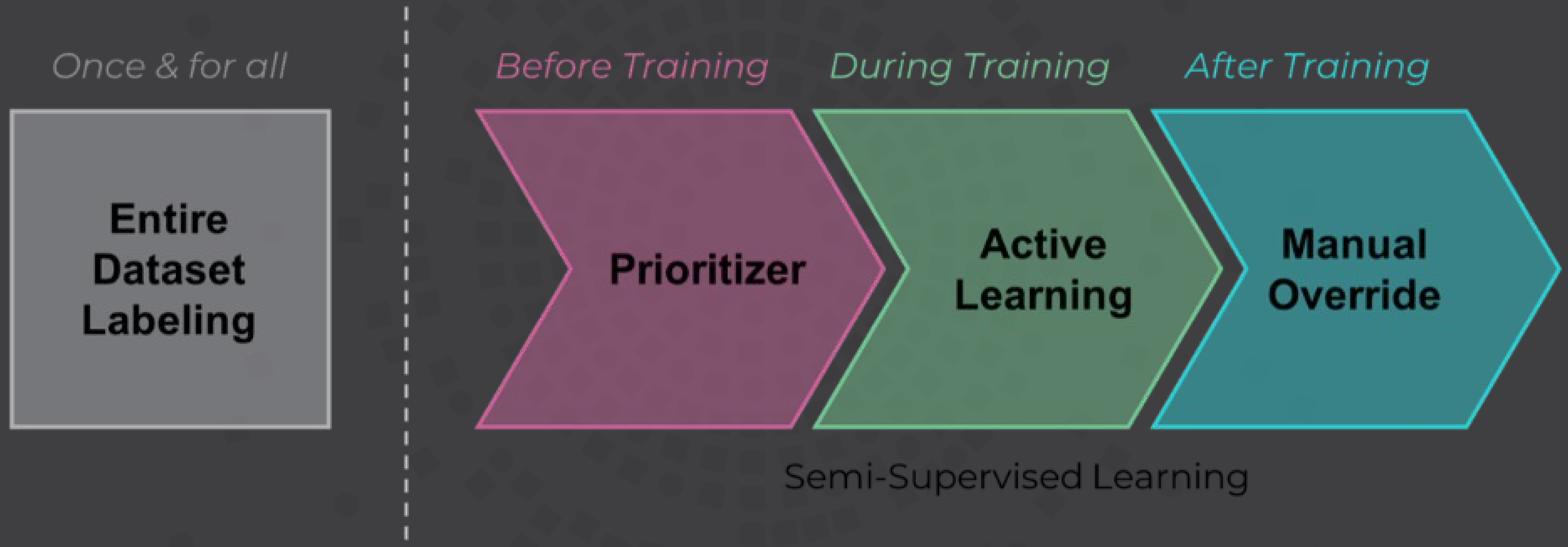
CNN

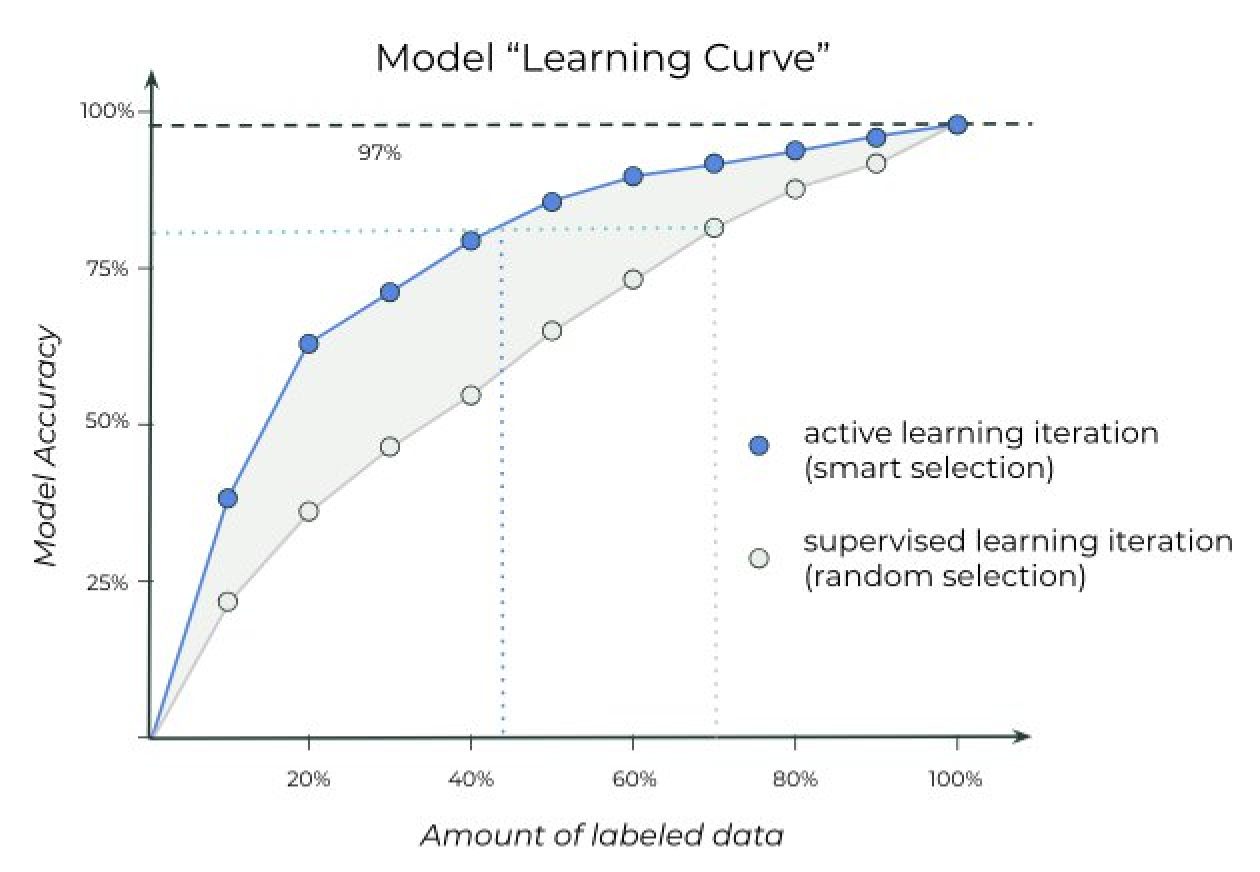
- Learns features
- Requires Large dataset of labeled images

Data is **abundant** but labeling is **expensive**




Cheap Labeling
Smart Labeling
Why Active Learning?
Model Convergence
Model Generalization


Once & for all
Entire Dataset Labeling
Before Training
Prioritizer
During Training
Active Learning
After Training
Manual Override
Semi-Supervised Learning


Model "Learning Curve"
100%
97%
75%
50%
25%
Model Accuracy
20%
40%
60%
80%
100%
Amount of labeled data
active learning iteration
(smart selection)
supervised learning iteration
(random selection)

Application to Parking Sign Recognition

Active Learning Framework

Can I park here?

Drivers spend a lot of time deciphering parking rules

- **Create traffic jams**
- **Endanger pedestrians safety**
- **Harm transportation environment**
- **High rate of parking tickets**

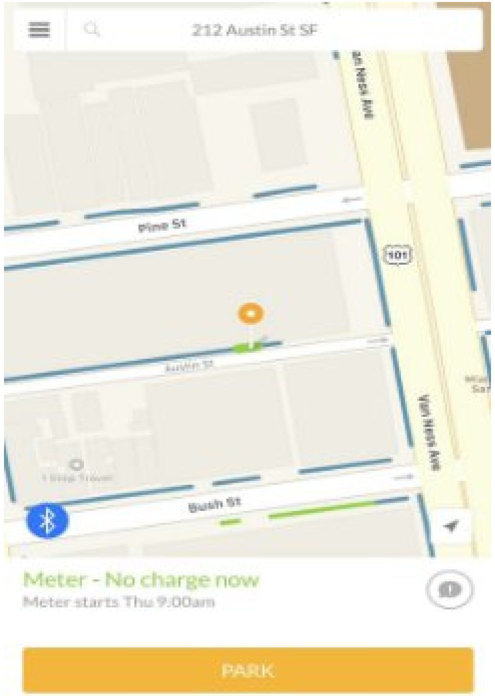
How computer vision can help to improve parking experience and transportation experience?

Structuring The Data

Online street-level imagery provides opportunities for developing new vision based algorithm

- Detect, classify and localize parking meters

Transferring the parking rules from images to maps



Data collection and structuring Is the first step to build any model

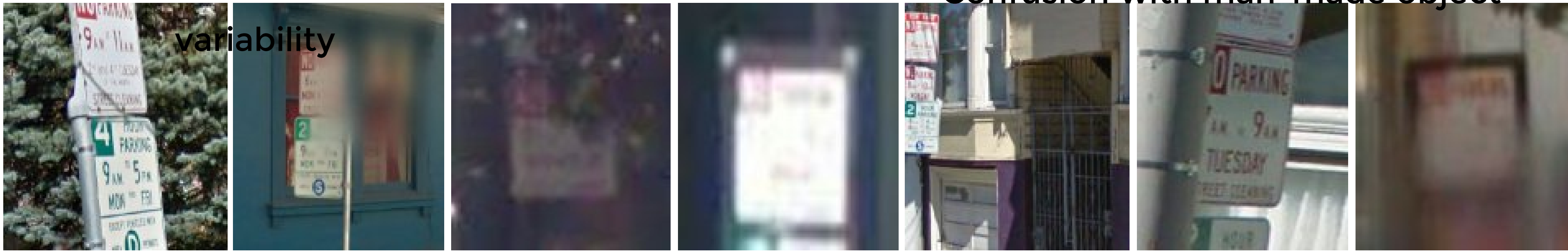
Street-level image collection and visualization

- Google street-view, Microsoft Streetside, Mapjack, EveryScape and …
- **Google Street-View**
    - 9 directional camera for 360 degree views at the height of 2.5 - 3 meters
    - multiple GPS units for positioning 3G/GSM/Wi-Fi antennas for scanning
    - 3G/GSM and Wi-Fi hotspots

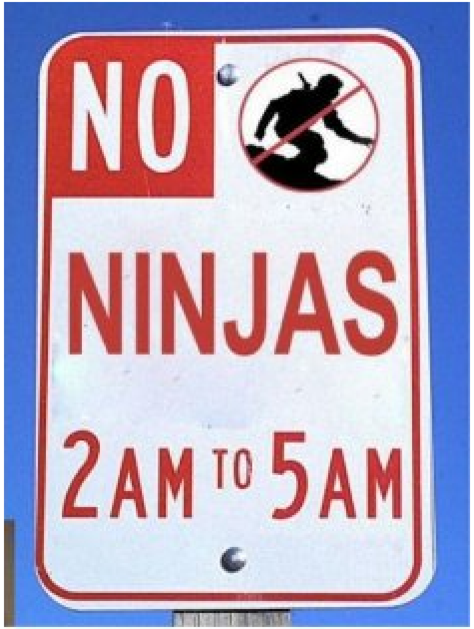
Developing computer vision models

It is a challenging task!

- **Appearance information**
  - different shapes, color and dimensions
  - contains a lot of text
  - Inter-class and intra-class variability
- **Standard computer vision challenges**
  - varying illumination
  - Pose and viewpoint
  - Occlusion
  - Confusion with man-made object



Making sense of a messy world



Data Collection and Annotation

Download and split panoramas into chips

Fine-tune the model & Identify where model is not performing well

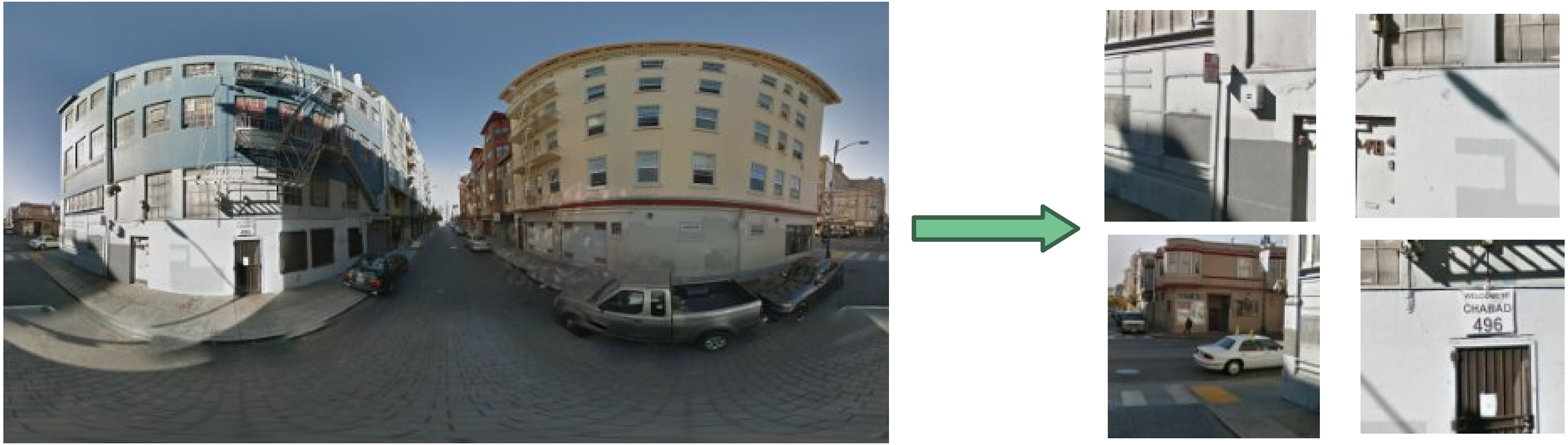
Label the images for Parking Sign

Human

Machine

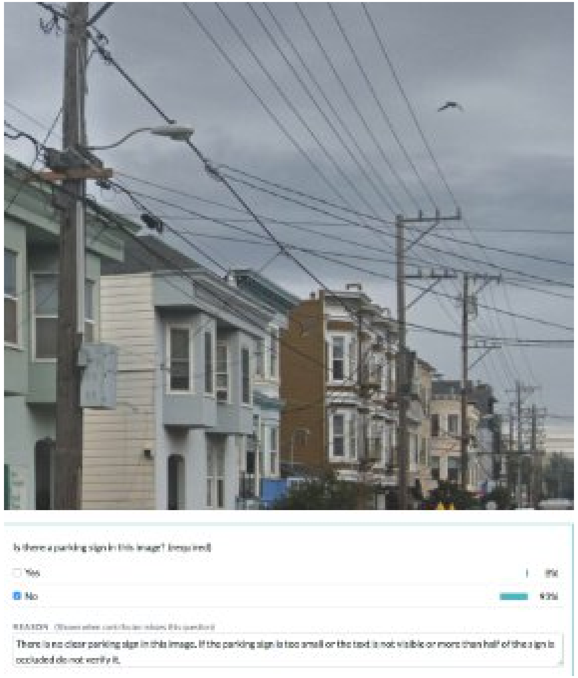
Data Collection

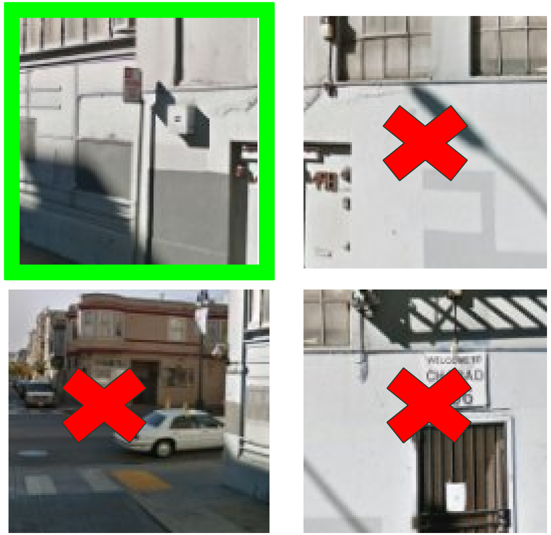
1. Download and Split Panoramas into Chips



Data Collection

2. Launch a Review Job to select Chips





Data Collection

3. Launch a Labeling Job to Box Parking Signs

Figure Eight Image Annotation Toolbox

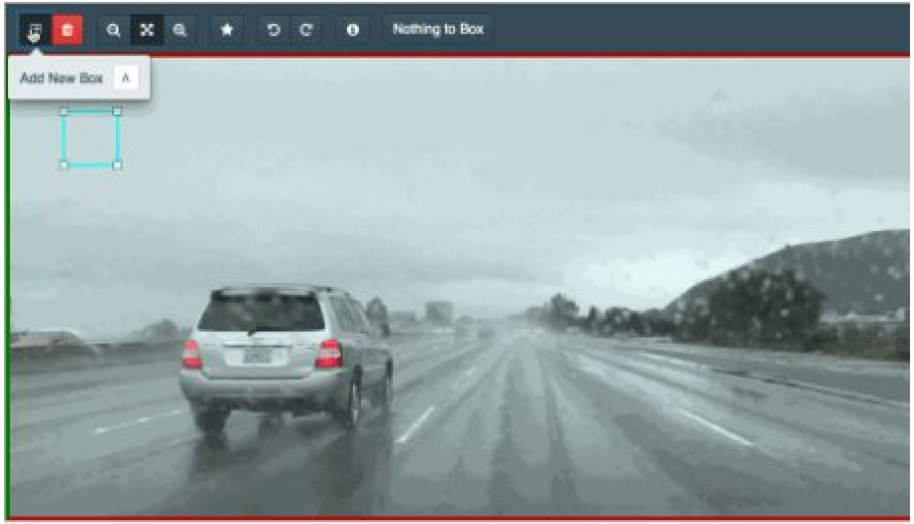


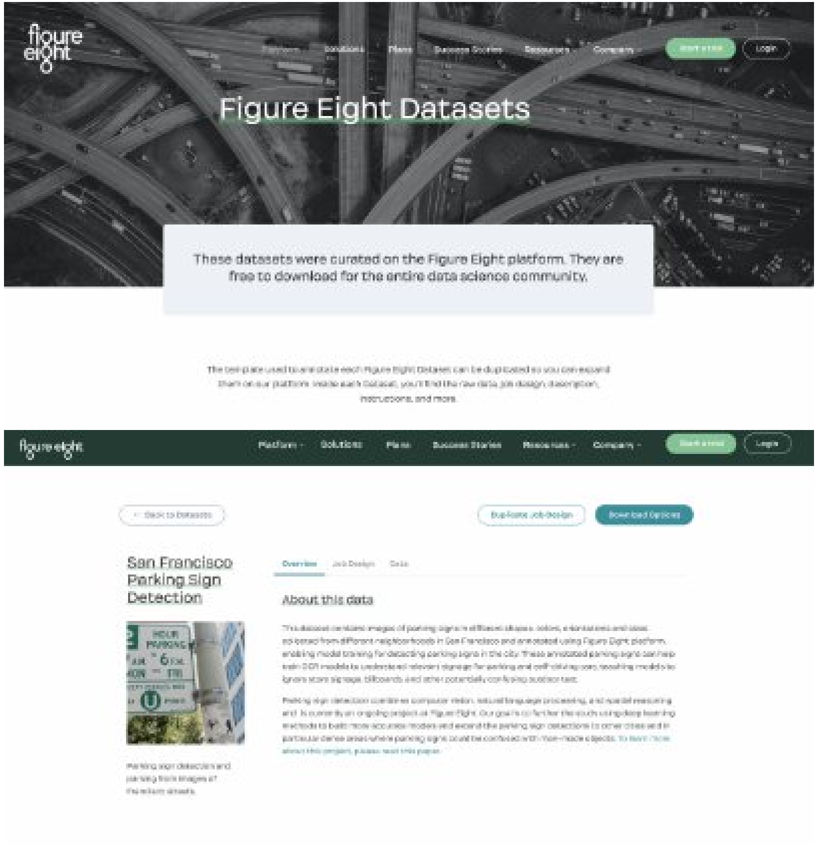
Figure Eight Human-In-The-Loop AI platform

Public Dataset

San Francisco street-level imagery

- Train Images
    - 1559 images
    - 2257 parking sign annotations
- Validation Images
    - 375 images
    - 606 parking sign annotations

www.figure-eight.com/datasets/



Building Deep Models for Parking Sign Recognition

YOLO vs SSD

Active Learning Approach used for Selection of Training Data

You Look Only Once (YOLO)

Used Darknet-19 classification model

- Mostly 3 x 3 filters
- Used batch normalization
- 19 convolutional layers & 5 maxpooling layers
- Initial trained on ImageNet (1000 categories)

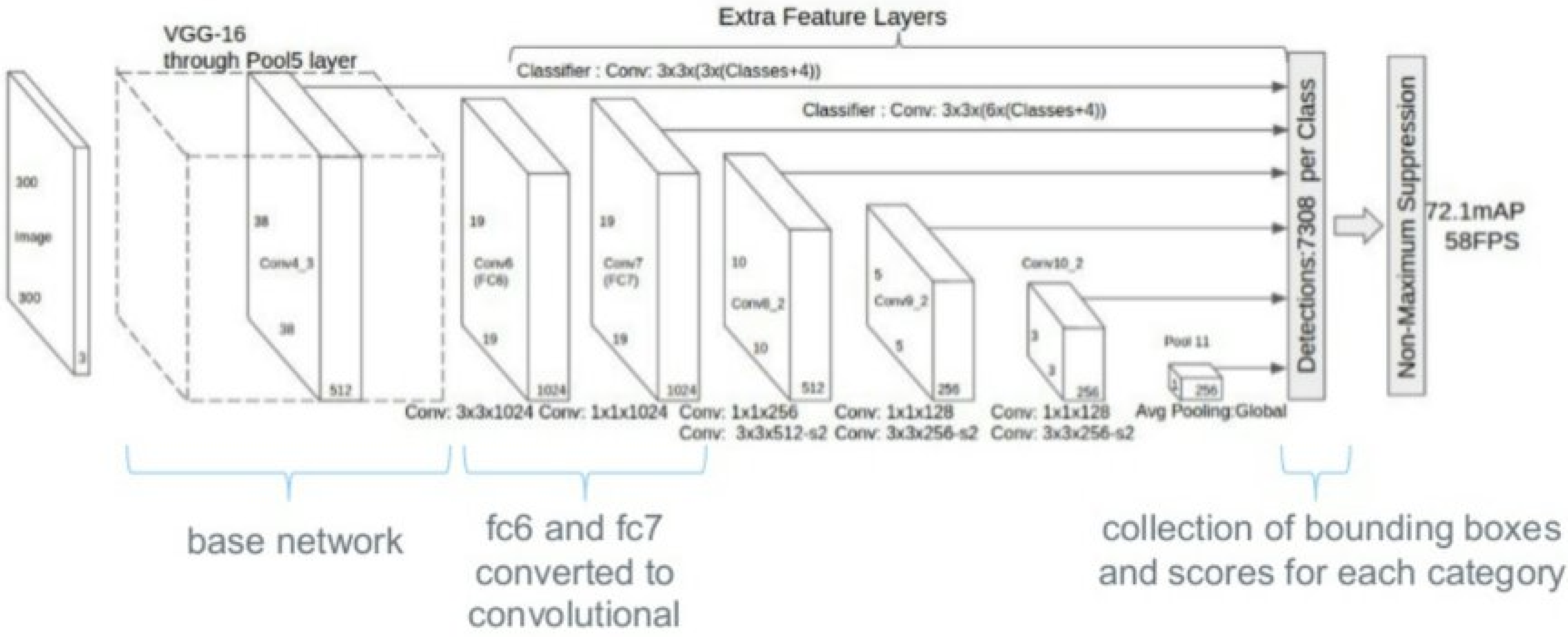
Type	Filters	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool		$2 \times 2/2$	112×112
Convolutional	64	3×3	112×112
Maxpool		$2 \times 2/2$	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		$2 \times 2/2$	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		$2 \times 2/2$	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		$2 \times 2/2$	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

Transfer Learning / Fine Tuning

- Remove last layer and replace with 3 x 3 convolutional layer with 1024 filters followed by 1 x 1 convolutional layer with the number of output
- Epoch is 160

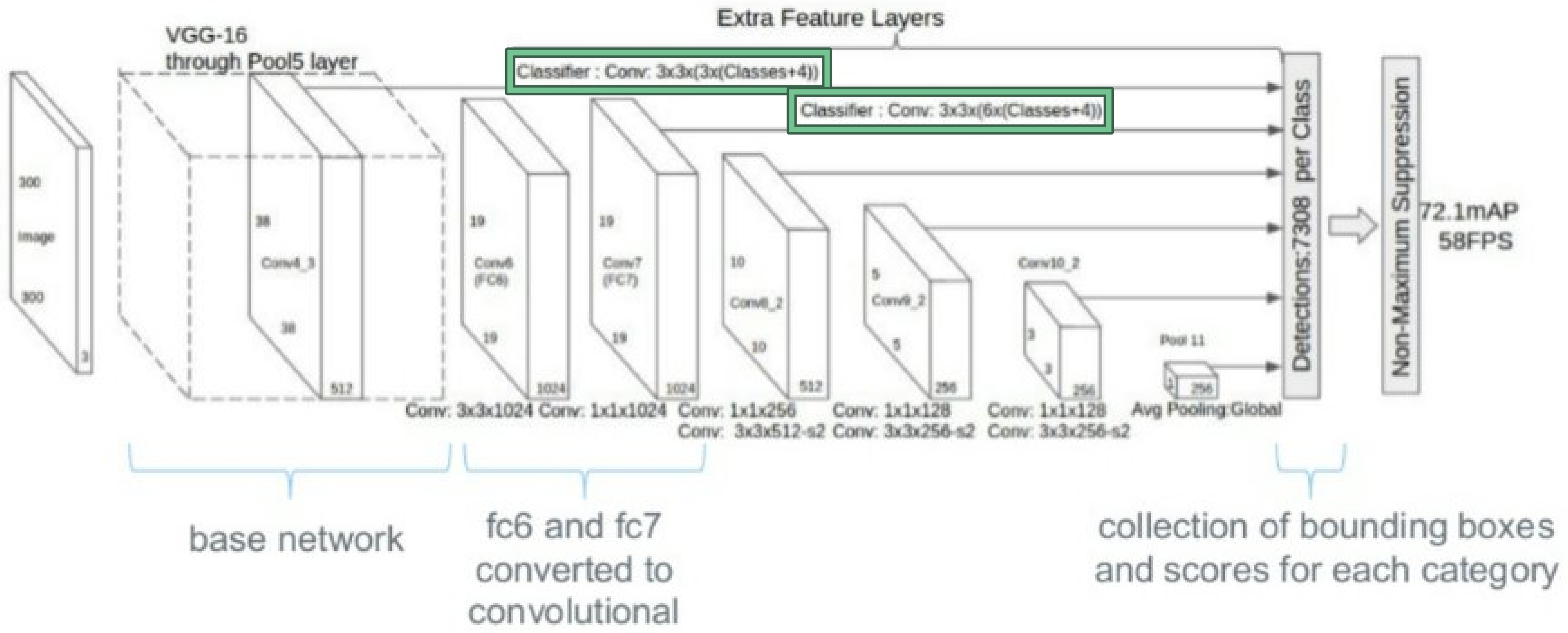
Single Shot Detector (SSD)

Multi-scale feature maps for detection



Single Shot Detector (SSD)

Apply on top of each conv feature map a set of filters that predict object with different aspect ratios and class categories



Active Learning Framework

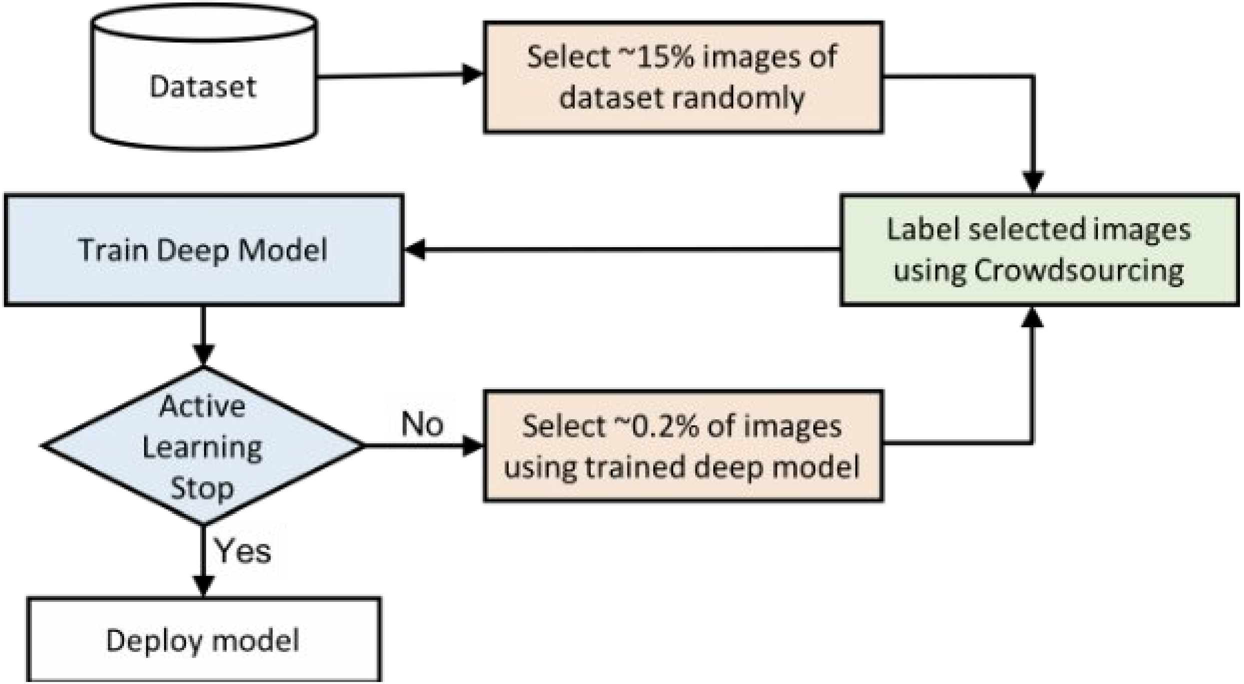
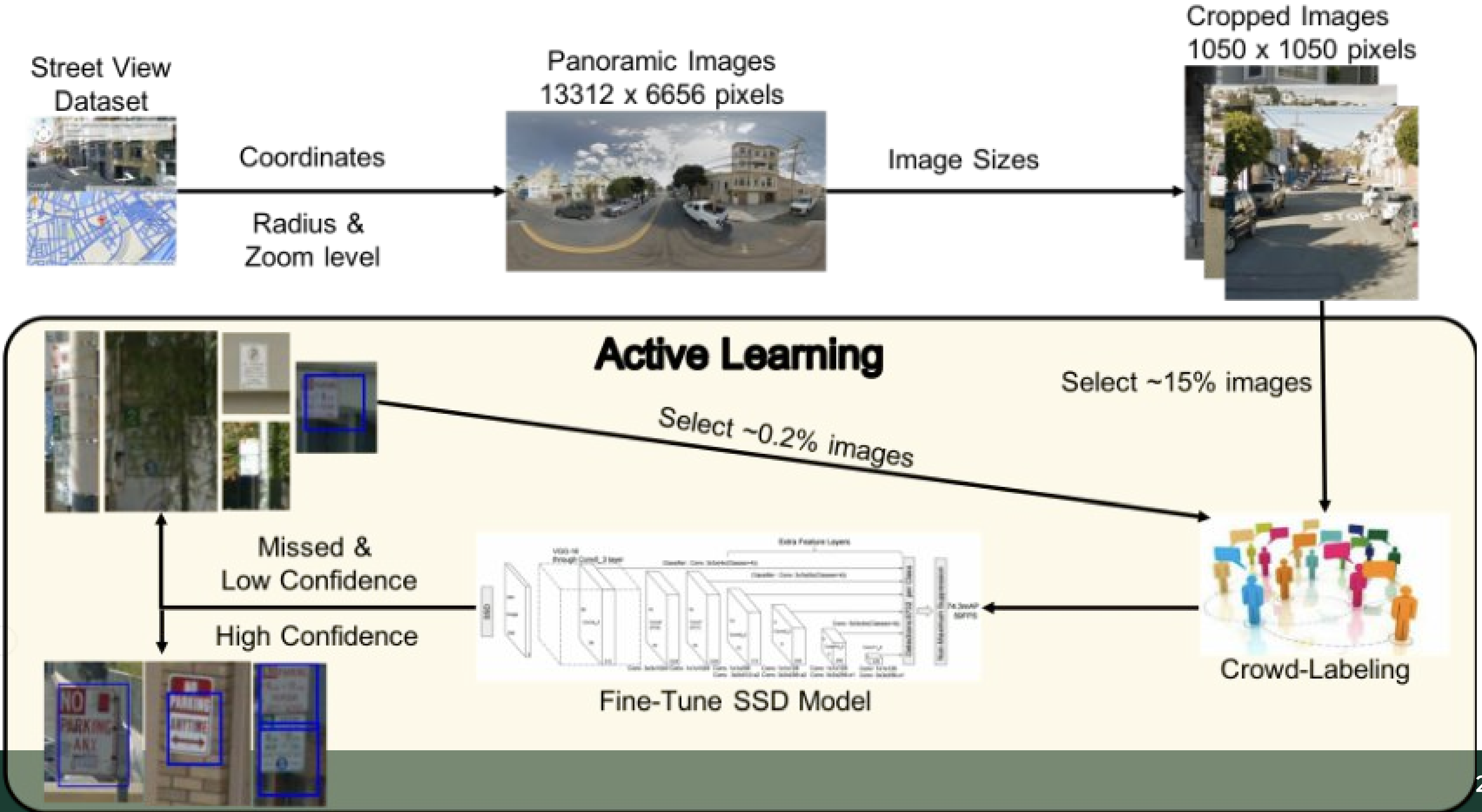


Image Selection and Crowding Labeling

Split Images into three subset and select images for labeling and training the model

- **High Confidence** images which have confidence above 80%
  - Select 20% images from the lowest confidence score
- **Low Confidence** images which have confidence below 80%
  - Select 60% images from highest confidence score
- **No Prediction** images which have no parking sign
  - Select 20% images randomly

Active Learning Framework with Object Detection



Training Sets

Selection of new images in training set using active learning framework

Dataset	No. of Images		No. of Annotations	
	New Addition	Total	New Addition	Total
Test Set	-	375	-	606
Training Set 1	509	509	704	704
Training Set 2	98	607	137	841
Training Set 3	380	987	589	1430
Training Set 4	550	1537	796	2226
Training Set 5	530	2067	893	3119
Training Set 6	400	2467	618	3737
Training Set 7	433	2900	707	4444

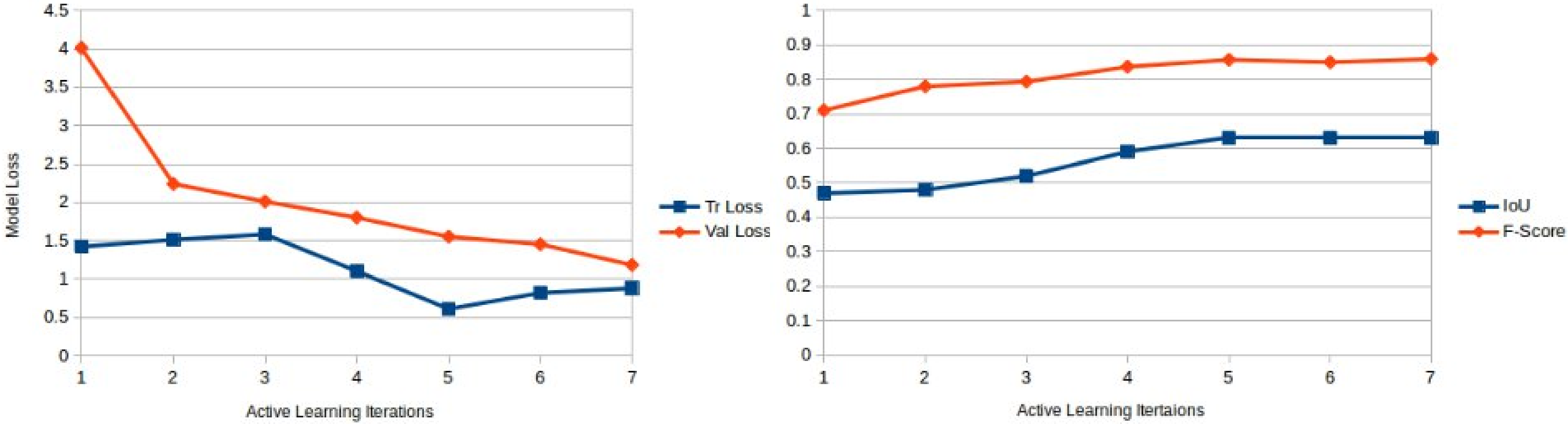
Results

Parking Sign detection results on test set after each iteration of Active Learning Framework

Active Learning Iterations	TP	FN	FP	Recall	Precision	F-Score	IoU
1	397	209	115	0.66	0.78	0.71	0.47
2	413	193	40	0.68	0.91	0.78	0.48
3	417	189	28	0.69	0.94	0.79	0.52
4	452	154	22	0.75	0.95	0.84	0.59
5	493	113	51	0.81	0.91	0.86	0.63
6	477	129	39	0.79	0.92	0.85	0.63
7	476	130	26	0.79	0.95	0.86	0.63

Active Learning Framework Performance

Decrease in model loss and increase in model accuracy



SSD Predictions

Challenging cases

How well does the model work on new data?

New York

Los angeles

Automated Parking Rules Extraction

"NO PARKING AM o12 NOON TUESDAY STREET CLEANING"

"2 HOUR PARKING TO MON THRU SAT EXCEPT VEHICLES WITH PERMITS AREA PARK AT 90 DEGREES"

Improving text analysis results through crowdsourcing

- **Detect text bounding boxes**
- **Extracting text for each box**
- **Add missing boxes and edit text by crowdsourcing**
- **Re-train text analysis model using new labeled textboxes**

Final Remarks

- Find challenging cases where system fails to accurately detect
- Reduce the redundancy in training data
- Save time and cost for labeling training data
- Improve the model training by better generalization

Thank You

figure-eight.com