Comparison of various CNN-based approaches for Crowd Counting

Abhishek Varma Dasaraju (2055979)

Manoj Kumar Nagabandi (2039097)

Supervised by Prof. Lamberto Ballan

Agenda

- Introduction
- Datasets
  - Mall Dataset
  - ShanghaiTech Part A and Part B
- Architectures
  - Object detection based
  - Density based
- Experiments
- Results
- Conclusion

Introduction

Why crowd-counting is needed ?

- Public Safety
  - Social distancing
  - Crowd management
  - Natural Disasters, fires
- Video Surveillance
  - Retail stores
  - Transportation hubs
  - Public places like stadiums and parks

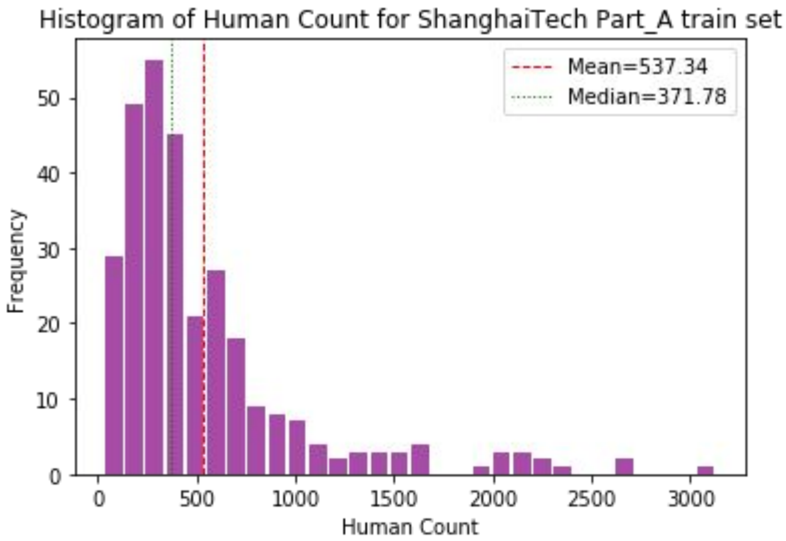
DATASETS

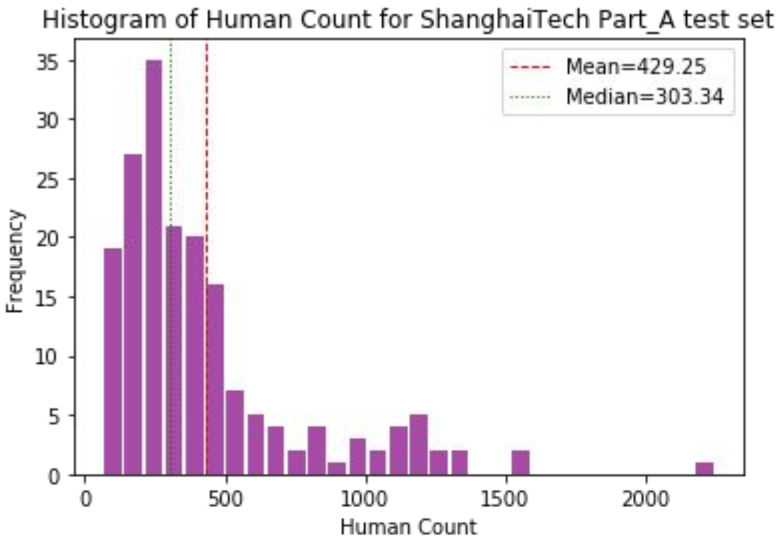
ShanghaiTech Part-A

- All images have **huge density crowds** with varied sizes.
- Part-A is subdivided into 300 train and 182 test images [1].
- Each image annotation have been obtained directly from the corresponding .mat file.

ShanghaiTech Part-A

- All images have **huge density crowds** with varied sizes.
- Part-A is subdivided into 300 train and 182 test images [1].
- Each image annotation have been obtained directly from the corresponding .mat file.



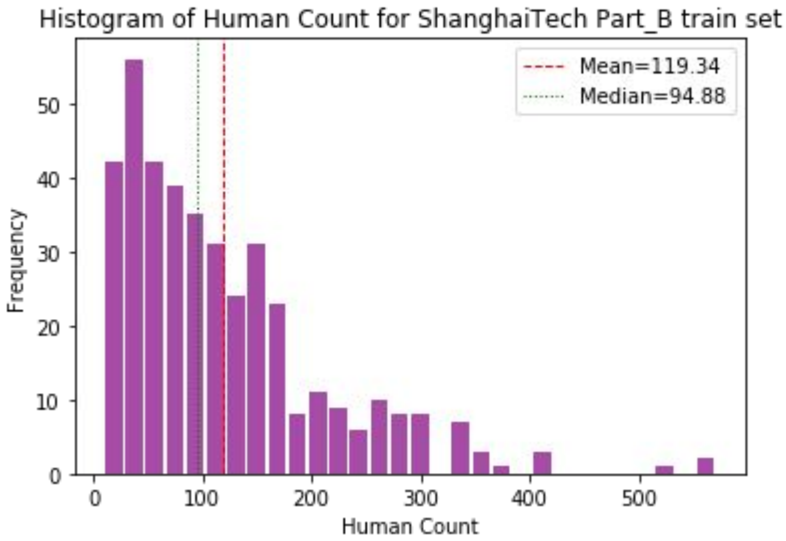


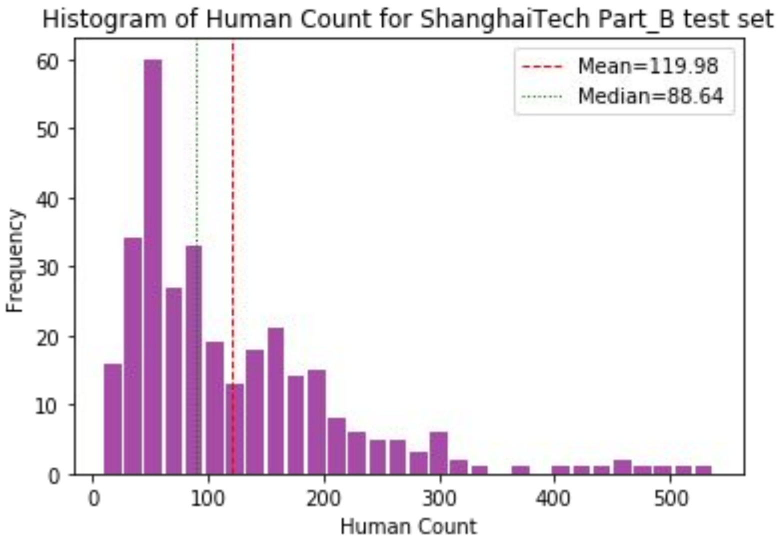
ShanghaiTech Part-B

- All images have **sparse crowds** with varied sizes.
- Part-B is subdivided into 400 train and 316 test images [1].

ShanghaiTech Part-B

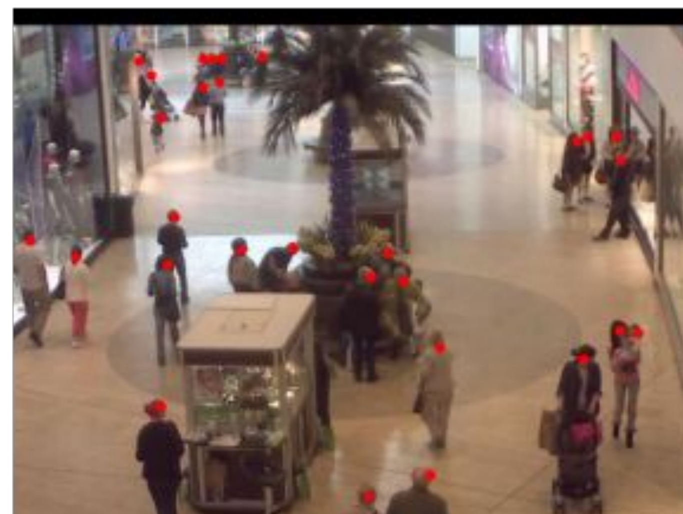
- All images have **sparse crowds** with varied sizes.
- Part-B is subdivided into 400 train and 316 test images [1].





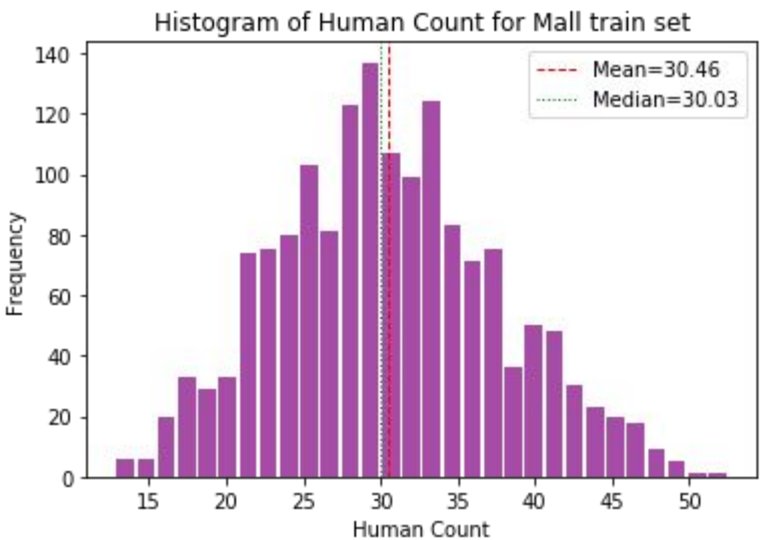
Mall Dataset

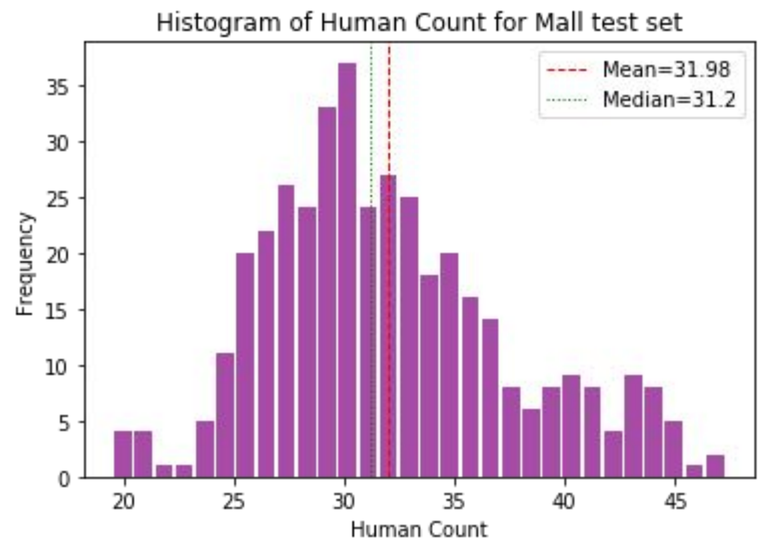
- There are a total of 2000 images with resolution of 480 x 640 each.
- Each image annotation have been obtained directly from the single .mat file.
- For training - 1600 images and testing - 400 images[1] are utilized.



Mall Dataset

- There are a total of 2000 images having same resolution of 480 x 640.
- The annotations are stored in a single .mat file for all of the images.
- For training - 1600 images and testing - 400 images are utilized [2].





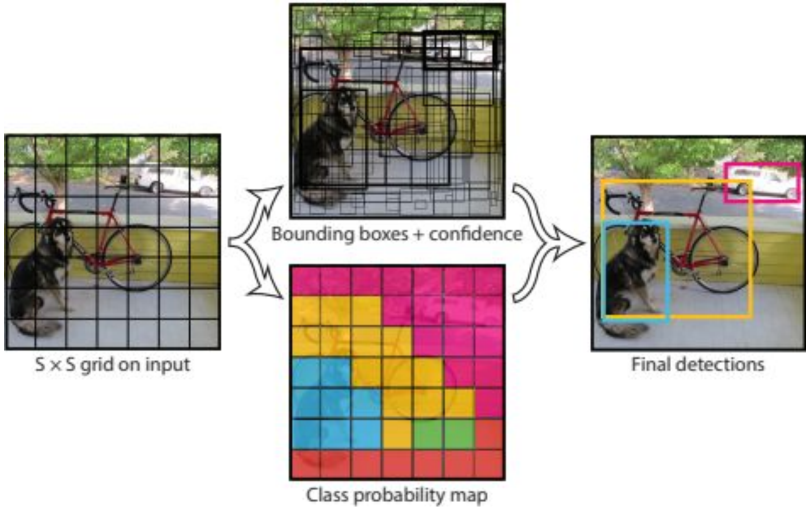
Architectures

- **Object detection based**
  - Yolo (V5 , V7, V8)
  - Faster R-CNN
  - SSD
  - EfficientDet

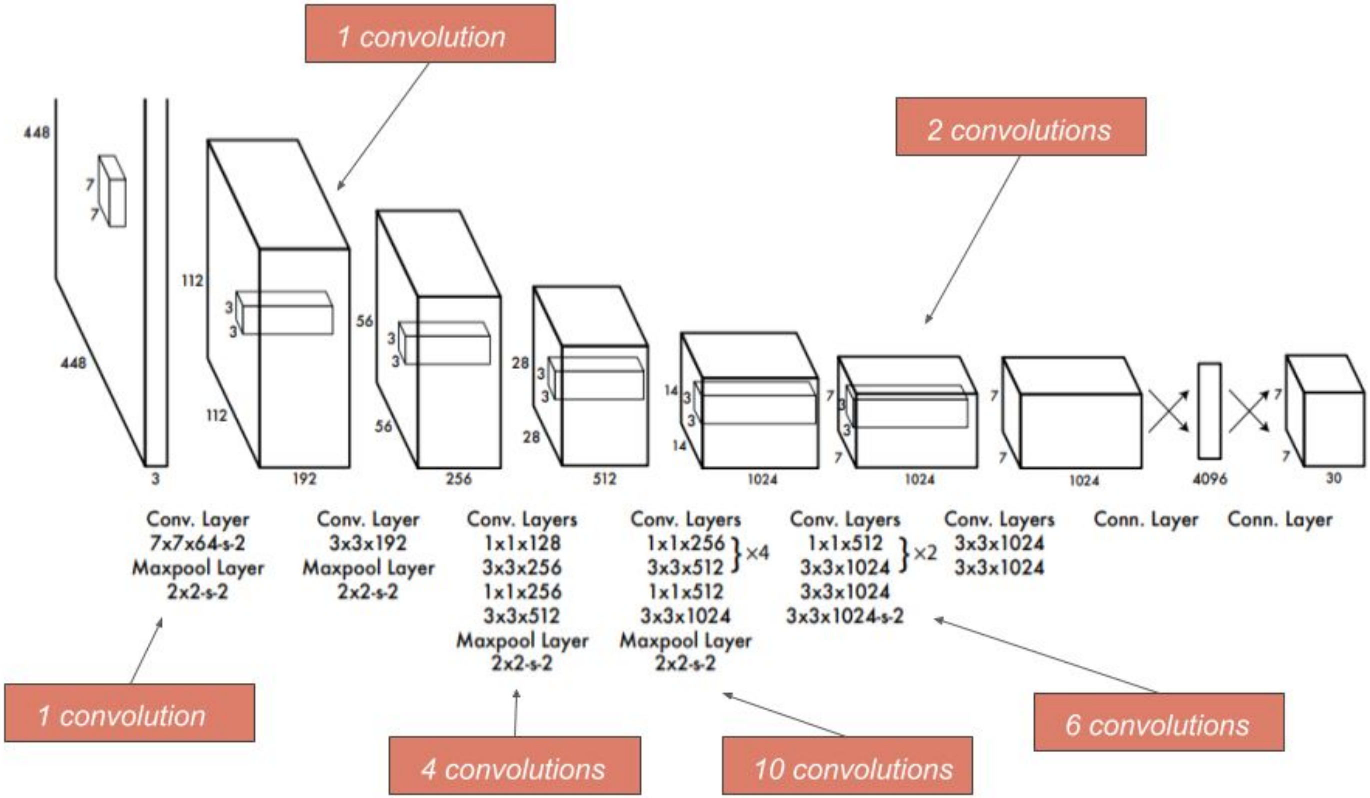
- **Density based**
  - MCNN
  - CSRNet

YOLO (You only look once)

- YOLO is an single CNN which simultaneously predicts multiple bounding boxes and class probabilities for those boxes.
- It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities.
- These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.



YOLOv1 Architecture



Yolo Differences

	YOLOv5 (2021) [3]	YOLOv7 (2022) [4]	YOLOv8 (2023) [5]
Backbone	CSPDarknet53	CSPDarkNet53	CSPDarkNet68
Input size	416x416,640x640, 1024x1024	416x416,640x640, 1024x1024	640x640,1280x1280,1536x 1536
Output stride	32	32	32
Neck	Spatial Pyramid Pooling layer(SPP)	Path Aggregation Network(PAN)	Path Aggregation Network(PAN)
Head	B x (5+ C) output layer	YOLO v5 head	Spatial Attention Module + YOLO v5 head
Loss Function	Focal loss	Focal, IoU, GIoU	Focal, IoU, GIoU, DIoU
Optimizer	SGD	Adam	Adam
Learning rate	0.002-0.01	0.001	0.002-0.0001

Sample test prediction of YOLO

YOLOv5x6: 17 persons

YOLOv7ex6: 25 persons

YOLOv8n: 29 persons

True person count: 35

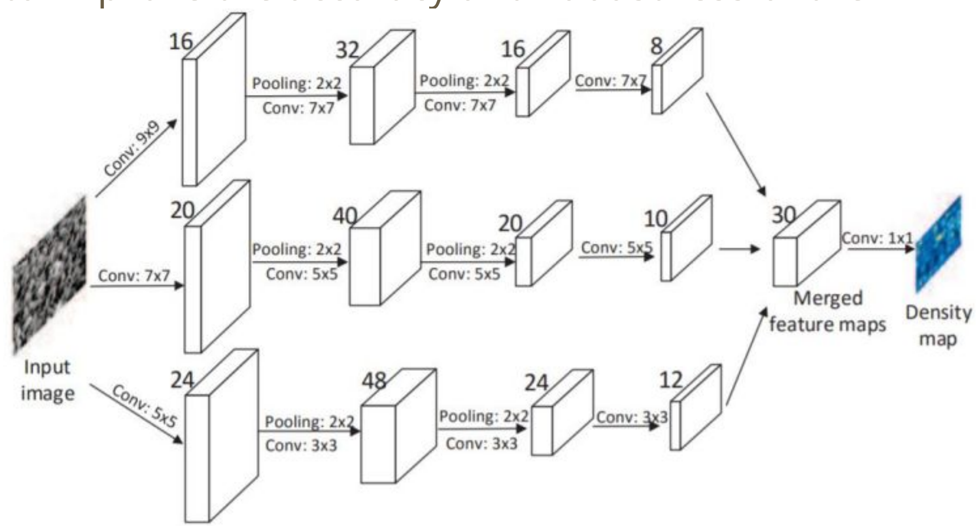
Density Map Generation

For density based approaches (i.e MCNN and CSRNet), ground truth density maps are generated for the images based on head annotations.The process is:

- K-dimensional tree is created using the non-zero elements of the ground truth density map.
- KD tree is queried to get the nearest neighbors for each point.
- Density is computed using a Gaussian filter.
- The sigma value for the Gaussian filter depends on the distances to the nearest neighbors.
- The computed density map is stored.

MCNN (Multi-column convolutional neural network)

- Multi-column convolutional neural network (MCNN) consists of multiple independent columns, each with filters of different scale, to capture both global and local information about the crowd [1].
- MCNN uses an ensemble approach to improve the accuracy and robustness of the network.
- Each column is trained on a different subset of data and the final output is a weighted combination of the predictions from all columns.
- MSE is used as loss function between ground truth density map and density map generated from the MCNN.



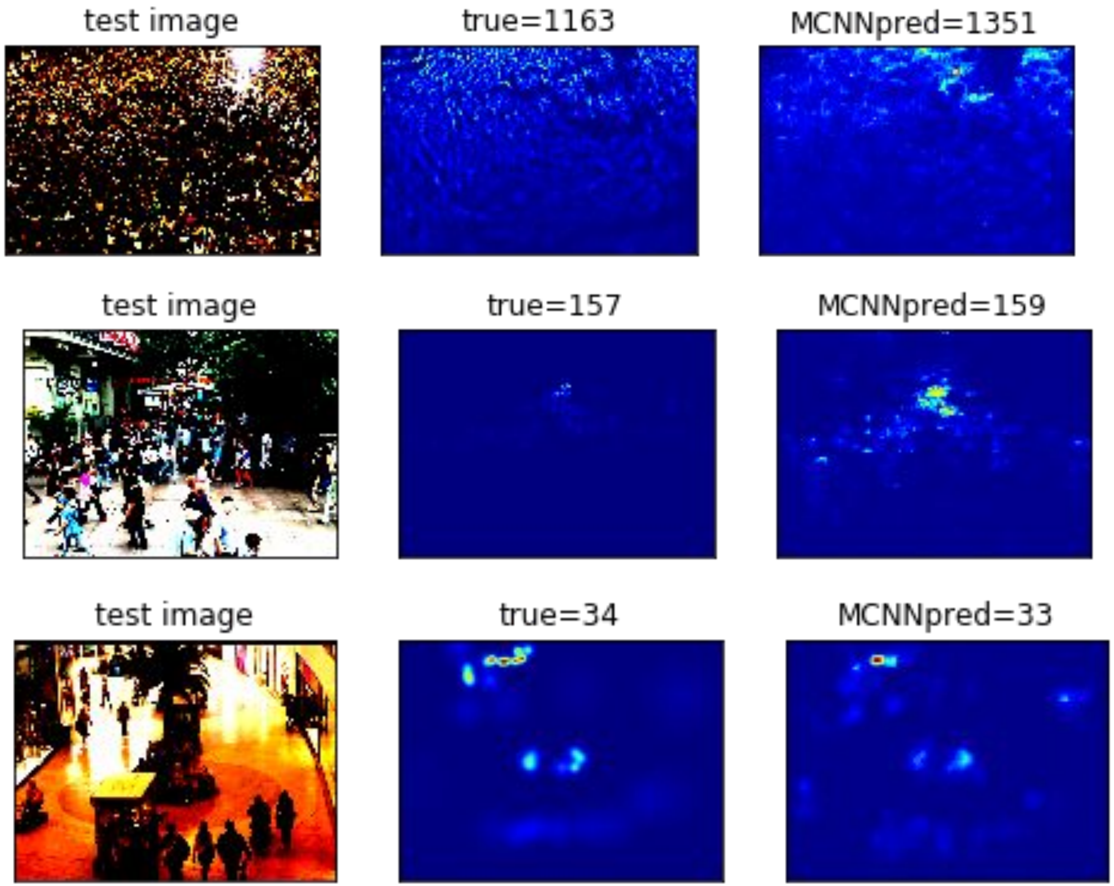
Training procedure

- MCNN architecture trained and tested on ShanghaiTech Part-A , Part-B and Mall dataset without any pre/processing of input image.

TRAINING RESULTS

Name of the Dataset	ShanghaiTech Part-A	ShanghaiTech Part-B	Mall Dataset
No of epochs Trained	350	150	92
Best Train MAE	158.31	19.96	3.01
Best Train MAE at epoch	348	149	54
Learning rate	1e-6	1e-6	1e-6
Batch size	1	1	32
Optimizer	Sgd with momentum	Sgd with momentum	Sgd with momentum

Test sample predictions



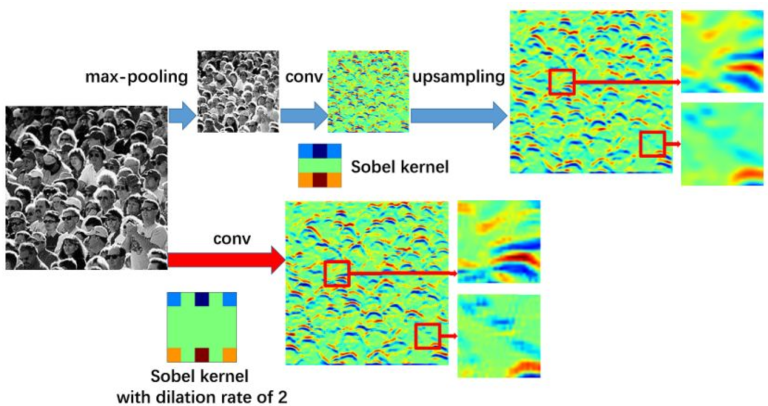
MCNN trained on ShanghaiTech Part-A

MCNN trained on ShanghaiTech Part-B

MCNN trained on Mall dataset

CSRNet

- It mainly comprises of front-end and back-end networks. [6]
- Front-end is basically, VGG-16 removing fully connected layers, leaving behind 13 layers.
- Back-end consists of 7 dilation convolution layers.
- The dilation rate of 2 yielded best results in previous experiments. Hence, this particular architecture.



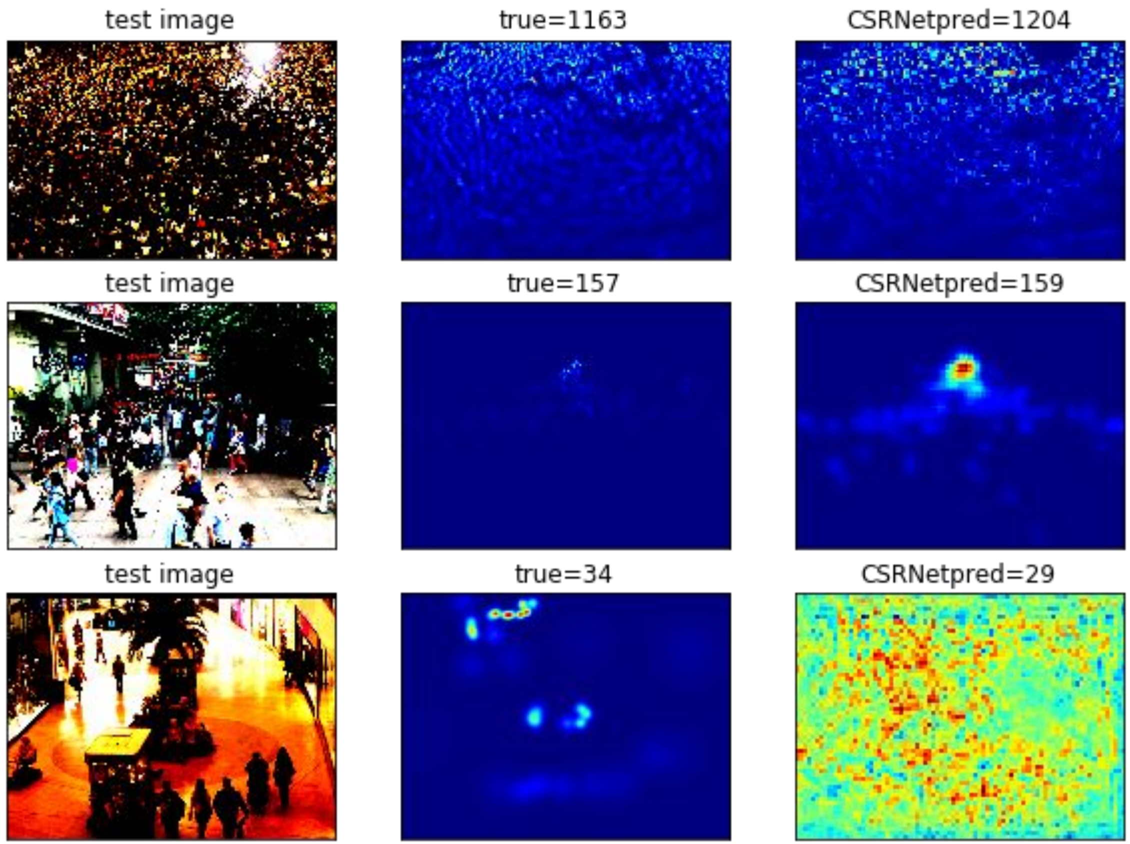
Configuration of CSRNet
input(unfixed-resolution color image)
front-end (fine-tuned from VGG-16)
conv3-64-1 conv3-64-1
max-pooling(stride=2)
conv3-128-1 conv3-128-1
max-pooling(stride=2)
conv3-256-1 conv3-256-1 conv3-256-1
max-pooling(stride=2)
conv3-512-1 conv3-512-1 conv3-512-1
back-end(dilation convolution layers)
conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-2 conv3-128-2 conv3-64-2
conv1-1-1

Training procedure

- CSRNet pre-trained on ShanghaiTech Part-A and Part-B tested for Mall dataset.
- It is also trained and tested on Mall dataset.
- The configuration of best model is kept constant[1].
- The only change is with batch size to achieve faster training.

Name of the Dataset	ShanghaiTech Part-A & Part-B	Mall Dataset
Transformations applied	Standard scaling across the channels	Standard scaling across the channels
No of epochs	200	50
Learning rate	1e-7	1e-7
Batch size	1	32
optimizer	Sgd with momentum	Sgd with momentum

Test sample predictions



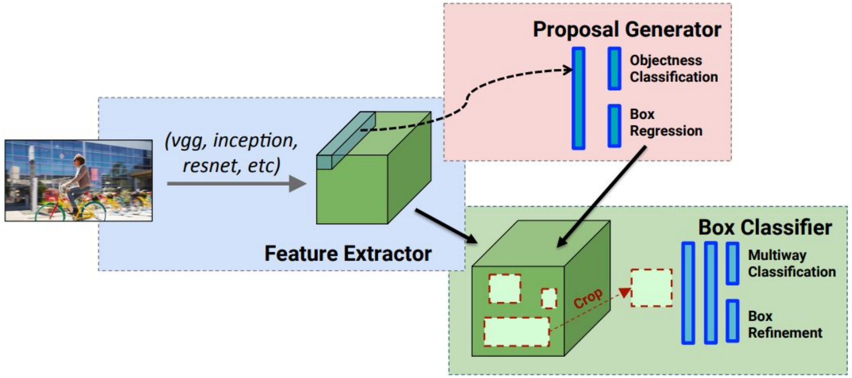
CSRNet trained on ShanghaiTech Part-A

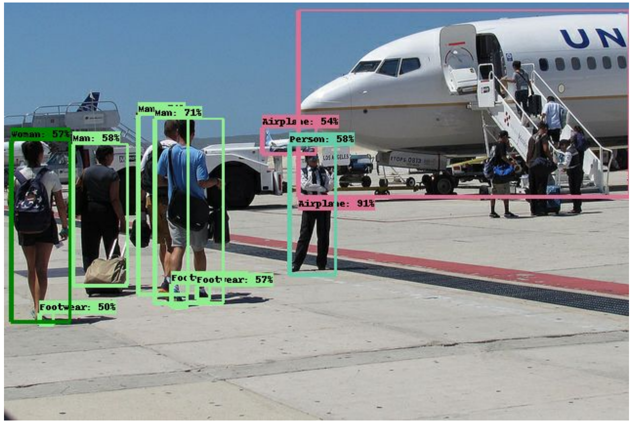
CSRNet trained on ShanghaiTech Part-B

CSRNet trained on Mall dataset

Faster R-CNN

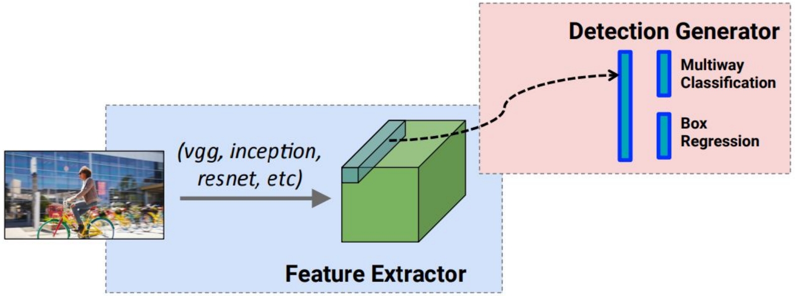
- Detection happens in two stages.
- Feature Extractor(Inception+ResNet V2) pre-trained on ImageNet.
- First stage, Region Proposal Network(RPN), predicts class agnostic box proposals.
- Second stage, predicts class and class-specific box refinement for each proposal.
- The Faster R-CNN with Inception+ResNet V2 feature extractor is fine-tuned on Open Images v4 dataset.
- The model can detect maximum 100 objects from 600 categories.

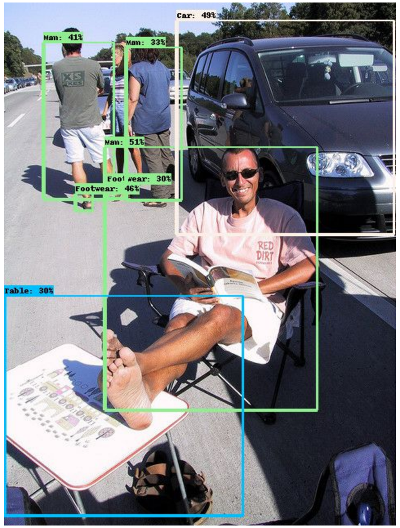




SSD

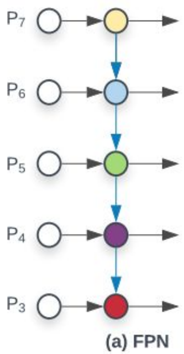
- In contrast to Faster R-CNN, SSD use a single feed-forward convolution network to directly predict classes and box encodings.
- Feature Extractor(MobileNet V2) pre-trained on ImageNet to extract features.
- SSD with MobileNet V2 is fine-tuned on Open Images V4 dataset as well.
- SSD capable of detecting as high as 100 objects present in the image.

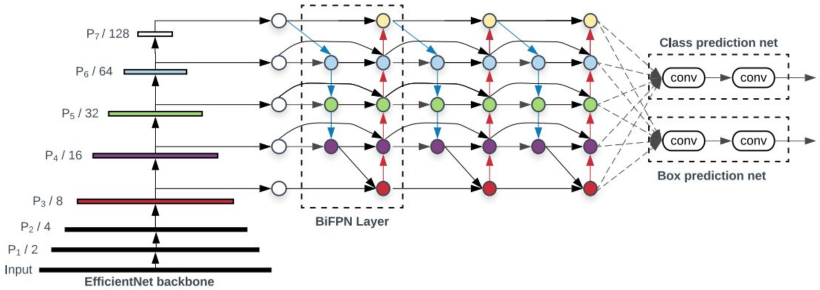




EfficientDet

- EfficientDet is also another one-stage detector.
- EfficientNet backbone network pre-trained on ImageNet gives out features at levels 3 to 7.
- These features undergo fusion in both directions with the help of BiFPN network.
- The output fused features extracted are fed to class and box predictions networks.
- This entire architecture is trained on COCO 2017 dataset.





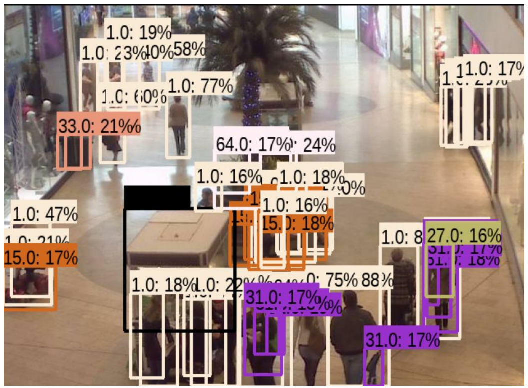
Experiment - object detection

- For object detection, Tensorflow hub object detection API’s are used to detect objects in the images.
- To test these approaches, last 400 images of Mall dataset is being used and counted the person classes detected in the image.

Observations:

- Faster R-CNN is having high accuracy but takes longer for inference.
- SSD is faster with inaccuracies.
- EfficientDet is accurate as well as efficient in terms of prediction time taken.

Sample test prediction of object detectors



EfficientDet: 31 persons

SSD: 52 persons

Faster R-CNN: 31 persons

True person count: 35

Density based methods results

Models	Train Dataset	Test dataset	MAE	Prior outcomes(MAE)
MCNN	PART - A PART - B PART - A PART - B MALL	PART - A PART - B MALL MALL MALL	133.96 22.53 19.77 24.04 2.74	110.2 [1] 26.4 [1] - - 3.15 [2]
CSRNet	PART - A PART - B PART - A PART - B MALL	PART - A PART - B MALL MALL MALL	65.92 11 11.07 9.28 4.57	68.02[1] 10.6[1] - - 3.15 [2]

Object Detection results

Models	MAE(20%)	Duration [s]
yolov5n6	26.37	2.16
yolov5s6	25.2	3.52
yolov5m6	23.63	6.32
yolov5l6	22.66	9.97
yolov5x6	21.94	17.51
yolov7e6e	4.2	76.99
yolov8n	6.3	37.76
FasterRCNN	4.15	564
SSD	9.97	72
EfficientDet	4.91	137

Summary of different approaches for crowd counting techniques

Category	Principles	Crowd Counting Accuracy	Location Accuracy	Annotation Complexity	Limitations
Detection-based	Detect then count; early approach	Low	High	High (object framing)	Low accuracy for highly crowded scenes
Regression-based	Directly learn to regress the count	Medium	N/A	Low (image-level count)	Less interpretable; lacks location information
Density map estimation	Compute number of people per pixel	High	Medium	Medium (head indication)	Low accuracy in low crowd scenes

CONCLUSION

- Yolov7 with e6e architecture shown superior performance in terms of MAE and speed. However, Faster R-CNN and EfficientDet achieved an MAE close to Yolov7e6e with more time for inference.
- MCNN achieved high performance with Mall Dataset but there are two significant drawbacks: long training time and ineffective branch structure.
- CSRNet dominated over MCNN in high crowd density situations.
- Density based approaches are effective when compared to object detection approaches due to their simplicity, accurate predictions and interpretable density maps.

References

- [1]. Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 589–597, 2016
- [2]. Ke Chen et al. Feature mining for localised crowd counting.
- [3]. Marko Horvat and Gordan Gledec. A comparative study of yolov5 models performance for image localization and classification. In Central European Conference on Information and Intelligent Systems, pages 349–356. Faculty of Organization and Informatics Varazdin, 2022
- [4]. Chien-Yao Wang, Alexey Bochkovskiy, and HongYuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696, 2022.
- [5]. Qiu Jing. Jocher Glenn, Chaurasia Ayush. Yolov8 by ultralytics. not yet published, Jan, 2023.
- [6]. Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1091–1100, 2018.