**Project Title:**
Deep Vision Crowd Monitor: AI for Density Estimation and Overcrowding Detection

**Document:**
Model Training (Part A & Part B)

**Student Name:** Sreeja Banoth
**Mentor:** G.K.S Jyoteesh

**Table of Contents**

# 1. Introduction

Crowd counting is a critical computer vision task used in video surveillance, public safety, traffic monitoring, and event management. Traditional object detection methods fail in dense crowd scenarios due to heavy occlusion. To address this, density map–based approaches are widely used.

In this project, **CSRNet (Convolutional Neural Network for Crowd Counting)** is implemented and trained on the **ShanghaiTech dataset (Part A and Part B)**. The model estimates crowd density maps, where the sum of pixel intensities corresponds to the total crowd count.

This document describes the **model training process and evaluation methodology,** for both dataset parts.

# 2. Dataset Description

### 2.1 ShanghaiTech Part A

- Contains highly congested crowd scenes
- Images collected from the internet
- High variation in density (extremely dense crowds)
- Ground truth provided as **point annotations (.mat files)**
- Considered challenging due to severe occlusion

### 2.2 ShanghaiTech Part B

- Low-density crowd images captured from street views
- More structured environments
- Easier compared to Part A
- Dataset split:
    - Training images
    - Testing images
- Ground truth provided as **point annotations**

# 3. Model Architecture – CSRNet

CSRNet is a density map–based convolutional neural network specifically designed for crowd counting.

3.1 Frontend (VGG-16 Backbone)

- Uses VGG-16 pretrained on ImageNet
- First convolutional layers are retained
- Acts as a feature extractor
- Pooling layers downsample feature maps to 1/8 resolution

3.2 Dilated Convolutions

- Uses dilated convolution layers
- Enlarges receptive field without loss of resolution
- Preserves spatial information
- Improves density estimation in dense regions

# 4. Training Methodology

## 4.1 Preprocessing

The following preprocessing steps were applied to both Part A and Part B:

- Images loaded in **RGB format**
- Resized to **512 × 512**
- Pixel values scaled to **[0,1]**
- ImageNet normalization applied (mandatory for VGG-16)
- Ground truth point annotations converted into **Gaussian density maps**
- Density maps downsampled by **8×**
- Density values multiplied by **64** to preserve total count
- Converted to PyTorch tensors

## 4.2 Training Configuration

- **Model:** CSRNet
- **Loss Function:** Mean Squared Error (MSE)
- **Optimizer:** Adam
- **Learning Rate:** 1e-5
- **Batch Size:** 1 (required for variable crowd scenes)
- **Epochs:**
    - Part A: Extended training
    - Part B: 20 epochs
- **Device:** CPU / Google Colab (T4 when available)

## 4.3 Evaluation Metrics

Model performance is evaluated using:

- **MAE (Mean Absolute Error)**
- **RMSE (Root Mean Squared Error)**

Lower values indicate better crowd estimation accuracy.

## 5. Experimental Results

5.1 Part A Results

- Dataset characteristic: High-density scenes
- Training stability improves over epochs
- Errors are higher due to extreme crowd congestion

Observations:

- CSRNet successfully learns dense crowd patterns
- Higher MAE/RMSE is expected for Part A

5.2 Part B Results

- Dataset characteristic: Sparse crowds
- Faster convergence compared to Part A
- Lower error values

Observations:

- Better crowd estimation accuracy
- Clean scenes lead to stable training
- MAE values are significantly lower than Part A

## 6. Conclusion

This milestone successfully completes:

- Training of CSRNet on **ShanghaiTech Part A & Part B**
- Implementation of complete preprocessing pipeline
- Evaluation using MAE and RMSE metrics
- Comparative performance analysis between dense and sparse datasets

The model demonstrates reliable crowd estimation and forms a strong foundation for final real-time inference and overcrowding detection.