# Crowd Counting in Frequency Domain

Rohan Guin

October 2024

## 1 Dataset

The ShanghaiTech dataset is a large-scale dataset for crowd counting, comprising 1,198 annotated crowd images. It is divided into two parts: Part A contains 482 images, and Part B includes 716 images. Part A is further split into 300 training images and 182 testing images, while Part B is divided into 400 training images and 316 testing images. Each individual in the crowd is marked with a point near the center of their head. In total, the dataset contains annotations for 330,165 people. The images in Part A were sourced from the Internet, while those in Part B were captured on the busy streets of Shanghai.

## 2 Implementation Details

The crowd counting framework used in this work operates in the frequency domain, which is a novel approach compared to traditional methods that focus on the spatial domain. The core idea is to transform the density map from the spatial domain into the frequency domain using characteristic functions, allowing for a more compact and organized representation of spatial information.

### 2.1 Model Architecture

The architecture consists of the VGG19 network for feature extraction, followed by a regression module made up of three convolutional layers. The key innovation in the model is the loss function, which operates in the frequency domain. The characteristic functions of both the ground-truth and predicted density maps are computed, and the L1 norm between these functions is used as the loss metric. By leveraging the properties of the characteristic function, this method avoids the need for external algorithms like the Sinkhorn or Hungarian algorithms, making the approach simpler and more efficient.

### 2.2 Training Process

The model is trained using the CHF loss function, which minimizes the discrepancy between the characteristic functions of the predicted and ground-truth density maps. The following configuration is used for training:

- **Optimizer:** Adam
- **Learning Rate:** $1 \times 10^{-5}$
- **Weight Decay:** $1 \times 10^{-4}$

Data augmentation techniques such as image resizing and cropping are employed during training, with specific crop sizes for different datasets:

- **ShanghaiTech A:** 128
- **ShanghaiTech B:** 512

.

## 3 Results

In the original study, five datasets were utilized: NWPU, JHU++, UCF-QNRF, ShanghaiTech A (SHTC A), and ShanghaiTech B (SHTC B). However, due to GPU limitations and to reduce the training and testing time, the experiments in this work were conducted specifically on the ShanghaiTech

dataset, focusing on its two subsets: ShanghaiTech A (SHTC A) and ShanghaiTech B (SHTC B).

Table 1: MSE and MAE values for SHTC A for ChfL model.

| | **MSE** | **MAE** |
|---|---|---|
| **SHTC A** | 102.3 | 70.5 |

The Mean Squared Error (MSE) of 102.3 reflects the average squared differences between predicted and actual crowd counts, with higher values indicating larger errors due to the squared penalty on deviations. This suggests there are some sizable errors in predictions, likely influenced by high variability in crowd density across test images, which is common in dense crowd counting tasks, especially in highly congested scenes. The Mean Absolute Error (MAE) of 70.5, on the other hand, provides a linear measure of error, indicating that on average, the model's predictions are off by about 70 people per image. While MAE does not penalize larger deviations as heavily as MSE, it gives a straightforward view of the model's performance by showing the average discrepancy per image. These metrics suggest a reasonable baseline but highlight room for further refinement, especially given the complex nature of crowd density in this dataset.