# Crowd Counting with Dual path Architecture Network

Rohan Guin

November 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

A dual-path network architecture for crowd counting can integrate both spatial and frequency domain features. Here is a detailed architecture proposal with dimensions for each layer. This design assumes a base input image size of 512x512 and aims to capture spatial information while performing frequency-domain analysis. The loss function I am using here is Mean Absolute Error (MAE). The whole architecture is shown in Table 1.

### 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.

This table describes a dual-path network architecture for crowd counting, consisting of two main processing paths: the spatial domain processing path and the frequency domain processing path.

The **spatial domain path** starts with an input RGB image of size $3 \times 512 \times 512$. This path involves several convolutional blocks, each of which applies convolution and pooling operations to progressively downsample and extract features from the image. In the first convolutional block, Conv1_1 and Conv1_2 layers with $64 \times 3 \times 3$ filters are applied, followed by a $2 \times 2$ max-pooling operation, reducing the spatial dimensions to $256 \times 256$. This process continues through subsequent convolutional blocks, increasing the number of filters in each block (from 64 to 512) while reducing the spatial dimensions, until reaching a size of $512 \times 32 \times 32$ in Conv4_2. At this stage, an Atrous Spatial Pyramid Pooling (ASPP) module applies three dilated convolutions with varying dilation rates (1, 2, 3), capturing multi-scale features, followed by concatenation and a $1 \times 1$ convolution to output a feature map of $512 \times 32 \times 32$. The decoder then gradually upsamples this feature map back to higher resolutions (up to $128 \times 128$), preparing the features for fusion with the frequency path.

The **frequency domain path** starts with the input image converted to the frequency domain using a 2D Fast Fourier Transform (FFT), which separates the image into magnitude and phase components of size $2 \times 512 \times 512$. Several convolutional blocks (Conv

Block Freq 1 to Conv Block Freq 3) then process the transformed image, reducing its spatial dimensions to $256 \times 64 \times 64$ through a series of convolutions and strided operations. The output is passed through a global average pooling layer to obtain a condensed feature representation of size $256 \times 1 \times 1$.

In the **fusion layer**, the feature maps from both paths are concatenated, resulting in a combined feature map of size $384 \times 128 \times 128$. This fused representation is then processed by two final convolutional layers, which reduce the feature dimensions to $64 \times 128 \times 128$. The output layer, a single-channel convolution, produces the final density map of size $1 \times 128 \times 128$, which provides the predicted crowd count across the image regions. This architecture combines both spatial and frequency information to improve accuracy in complex crowd counting tasks.

### 2.2 Training Process

The model is trained using Minimum Absolute Error which is used to measure the accuracy of predictions in regression tasks. It represents the average absolute difference between predicted and actual values, giving insight into the model's prediction accuracy without being overly affected by large errors. The following configuration is used for training:

- **Optimizer:** Adam
- **Learning Rate:** $1 \times 10^{-5}$
- **Epoch:** 50
- **Batch Size:** 8

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

- **ShanghaiTech A:** 128

.

## 3 Results

In the original study [1], 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 A (SHTC A) dataset.

Table 1: MSE and MAE values for SHTC A for Dual Path Architecture model.

| | **MSE** | **MAE** |
|---|---|---|
| **SHTC A** | 206.7 | 45.3 |

## 4 Reference

[1] Shu, W., Wan, J., Tan, K. C., Kwong, S., & Chan, A. B. (2022). Crowd Counting in the Frequency Domain. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.