CrowdDiff: Multi-hypothesis Crowd Density Estimation using Diffusion Models

Aravind Seshadri 210180

October 22, 2024

1 Introduction

Crowd counting is essential in surveillance, public safety, and crowd control. It is typically addressed using two methods: **Localization-based methods** and **Density-based methods**. The paper proposes using diffusion models for generating the crowd density maps.

2 Implementation

2.1 Pre-Processing

Each dataset is pre-processed to form multiple samples of an image. These samples are overlapping crops of a single image. The images' ground truth values or annotations are used to convolve the point information with a narrow Gaussian kernel, resulting in a density map. The density map is then denoised, and a binary threshold is applied. The resulting map is passed through a contour counter, which returns the number of circles or the crowd count in the image. The following images are samples from pre-processing an image from the shtech crowd dataset.

Original Image

Crop 1

Crop 2

2.2 Training

The model is trained using the dataset and an additional counting branch. The model is initialized with the ImageNet weights and trained to predict the ground truth density map. Since the final trained weights were available, this process has been **skipped** during implementation.

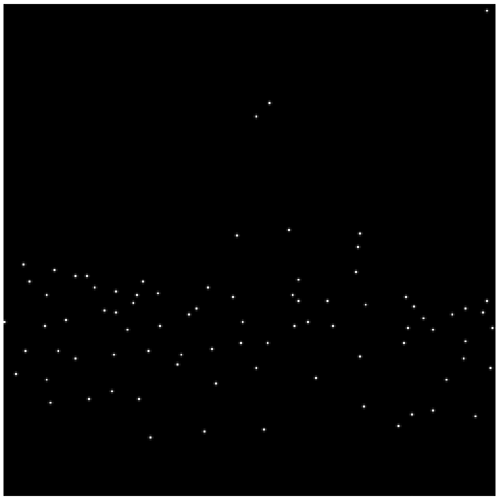
2.3 Testing

After pre-processing the data, the crowd maps are generated using a diffusion model. The implementation includes the following steps:

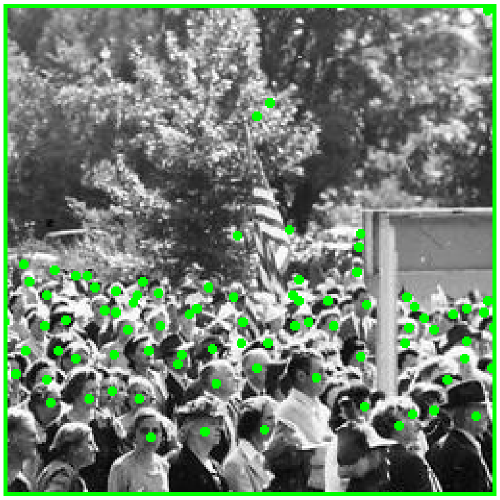
- Each image in the processed folder is loaded using a custom dataloader. The data loader stores the information on the actual crowd count from the ground truth density map. Additionally, each sample is stored in the data loader as a low-resolution image. The diffusion process uses this information to generate the high-resolution density map of that image.
- Using the low-resolution information, the guided diffusion process occurs for 1000 timesteps. The resulting density map is stored in the pred_density.png for an image.

- To count the predicted number of people in a density map, the obtained image is first passed through a denoising function. This is to smoothen the image. The image is then thresholded to form a binary image with white circles and a black background. A morphological operation of erosion followed by dilation is performed. Finally, the contours in the image are obtained, and the total number of contours gives us the crowd count for that image. A sample image with merged contours is shown below.

Test Image



Predicted Density Map



Combined image

2.4 Datasets

Jhu_Crowd ++	**Shtech_A**	**Shtech_B**	**UCF_CC_50**	**UCF_QNRF**
Contains 4372 annotated images with varying crowd densities and weather conditions.	Contains 482 images with images collected from the internet	Contains 716 images collected from the streets of shangai	Contains 50 images of extremely crowded regions collected from flickr.	Contains 1535 images that are very diverse, collected from the web.
Format of the dataset is test folder with images, gt(ground truth) and image_labels.	Format of the dataset is a folder test_data that contains images and ground-truths.	Format of the dataset is a folder test_data that contains images and ground-truths.	Format of the dataset is a test folder with img_no.jpg and img_no_ann.mat files.	Format of the dataset is a test folder with img_no.jpg and img_no_ann.mat files.

2.5 Evaluation

For evaluation, MAE and MSE are the metrics used. Since it is unclear how the paper's authors have combined the crops for each image, we take the average MSE and MAE for every image sample. This results in the final MSE and MAE. This value is calculated for the five benchmarks mentioned in the paper and the results are tabulated. **Note:** Due to limited resources on the GPU, the entire dataset was not processed, and only a subset was used to obtain the results. Hence there might be a variation in the results obtained compared to the paper.

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N} \|c_n - \bar{c}_n\|_1$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \|c_n - \bar{c}_n\|_2^2}$$

Metrics	**JHU-Crowd++**	**Shtech-A**	**Shtech-B**	**UCF_CC**	**UCF-QNRF**
MAE	51.82	46.352	6.85	128.202	77.4
MSE	203.194	81.25	12.2	261.28	117.103

Table 1: Implemented values for 5 benchmarks