Deep Learning for Segmentation and Crowd Counting: U-Net and MCNN

Fouzia Zilani

July 26, 2025

1 Introduction

Deep learning has revolutionized computer vision tasks such as image segmentation and crowd counting. U-Net, originally designed for biomedical segmentation, has been adapted for various tasks, including density map regression for crowd counting. MCNN (Multi-Column Convolutional Neural Network) is a specialized architecture for crowd counting, designed to handle scale variations in dense crowds. This report details the training, evaluation, and comparison of these models on public datasets.

2 Task 1: U-Net for Image Segmentation

2.1 Dataset

We use the **Dog Segmentation Dataset** (`https://www.kaggle.com/datasets/santhoshkumarv/dog-segmentation-dataset`), which contains images of dogs and corresponding binary masks.

2.2 Method

A minimal U-Net architecture was implemented in TensorFlow/Keras. Images and masks were resized to 128×128 and normalized. The model was trained using binary cross-entropy loss and Adam optimizer.

2.3 Results

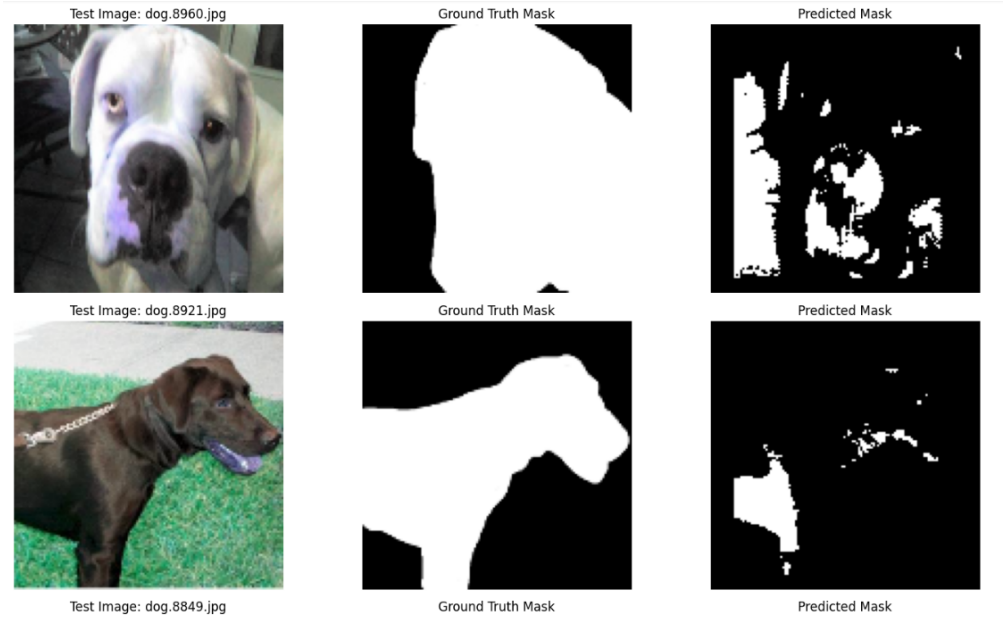

Figure 1: Example segmentation results: (a) Input image, (b) Ground truth mask, (c) Predicted mask.

2.4 Discussion

U-Net achieved high segmentation accuracy, effectively delineating dog boundaries. The skip connections in U-Net help preserve spatial information, which is crucial for segmentation tasks.

3 Task 2: U-Net for Crowd Counting

3.1 Dataset

We use the **ShanghaiTech Part A** dataset (`https://www.kaggle.com/datasets/tthien/shanghaitech`), which contains images of crowded scenes with point annotations.

3.2 Method

U-Net was adapted to regress density maps. The ground truth density maps were generated by applying Gaussian kernels to annotated points. The model was trained with mean squared error (MSE) loss.

3.3 Results

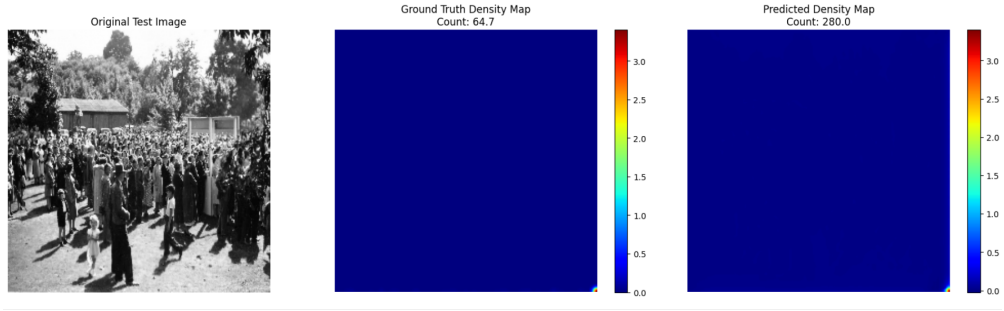

Figure 2: U-Net crowd counting: (a) Input image, (b) Ground truth density map, (c) Predicted density map.

3.4 Discussion

While U-Net can be adapted for crowd counting, its performance is limited compared to specialized architectures due to its generic design.

4 Task 3: MCNN for Crowd Counting

4.1 Dataset

The same ShanghaiTech Part A dataset was used.

4.2 Method

MCNN consists of three parallel convolutional columns with different receptive fields to handle scale variation. The outputs are concatenated and passed through a 1×1 convolution to produce the density map.

4.3 Results

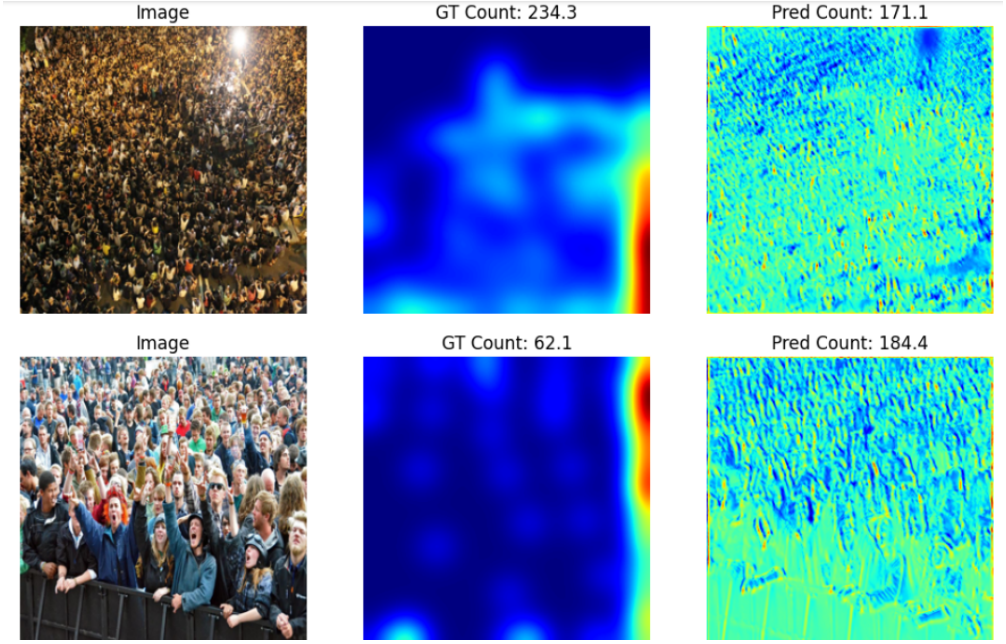

Figure 3: MCNN crowd counting: (a) Input image, (b) Ground truth density map, (c) Predicted density map.

4.4 Discussion

MCNN outperformed U-Net in crowd counting, thanks to its multi-column design that captures crowd density at multiple scales.

5 Task 4: Comparison of U-Net and MCNN for Crowd Counting

5.1 Loss Curves

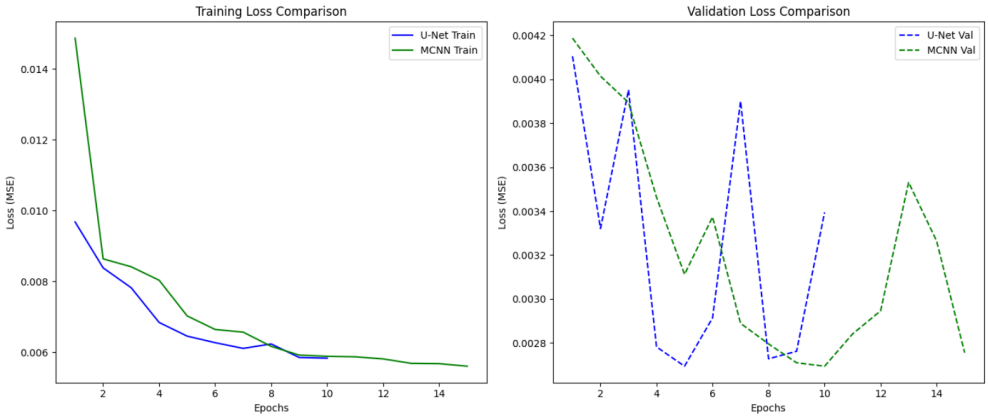

Figure 4: Training and validation loss curves for U-Net and MCNN.

5.2 Discussion

MCNN achieves lower MAE and MSE, indicating better counting accuracy. Its multi-scale columns are better suited for the scale variation in crowd images, while U-Net is more generic and flexible for segmentation tasks.

6 Conclusion

U-Net is effective for segmentation and can be adapted for crowd counting, but MCNN is superior for crowd counting tasks due to its specialized architecture. For segmentation, U-Net remains a strong choice.

Code is in this link: Link