**A Dilated Convolutional Neural Network for Cross-Layers of Contextual Information for Congested Crowd Counting**

[Zhiqiang Zhao](https://pubmed.ncbi.nlm.nih.gov/?term=%22Zhao%20Z%22%5BAuthor%5D), [Peihong Ma](https://pubmed.ncbi.nlm.nih.gov/?term=%22Ma%20P%22%5BAuthor%5D), [Meng Jia](https://pubmed.ncbi.nlm.nih.gov/?term=%22Jia%20M%22%5BAuthor%5D), [Xiaofan Wang](https://pubmed.ncbi.nlm.nih.gov/?term=%22Wang%20X%22%5BAuthor%5D), [Xinhong Hei](https://pubmed.ncbi.nlm.nih.gov/?term=%22Hei%20X%22%5BAuthor%5D), 2024

Methods:

The authors describe some kinds of methods:

* Multicolumn CNN-Based Methods
* Dilated Convolution-Based Methods

They propose Cross-Level Dilated Convolutional Neural Network (CL-DCNN). Method is used mainly for crowd counting tasks. The network comprises three key components: the backbone module, Dilated Contextual Modules (DCMs) and a fully connected layer. They use density map and gave the loss function formula. They also use two another metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Datasets:

The authors performed the experiment on several datasets:

* ShanghaiTech - stands as one of the most extensively utilized large-scale crowd counting datasets. It comprises 1198 images and 330,165 annotated head centers. The dataset is divided into two parts, part A and part B, reflecting distinct crowd density distributions. Part A consists of images randomly sourced from the Internet, whereas Part B comprises images captured from a bustling street in the metropolis of Shanghai.
* Mall - is a dataset compiled from surveillance videos of shopping centers. The video sequence encompasses 2000 frames and includes a total of 62,325 pedestrians. This dataset encapsulates a broad spectrum of density variations under various lighting conditions, along with instances of severe occlusion between individuals.
* UCF\_CC\_50 - developed at the University of Central Florida, comprises only 50 images with 63,075 annotated head centers. This dataset encompasses a diverse array of scenes, including concerts, protests, stadiums, and marathons, with densities ranging from 94 to 4543. The images in UCF\_CC\_50 exhibit different viewing angles, resulting in varying degrees of perspective distortion.
* UCF-QNRF - stands as a highly challenging dataset, comprising 1535 high-resolution crowd images with a total of 1,251,642 annotated head centers. These images encapsulate a wide spectrum of crowd densities and are captured by surveillance cameras with varying viewpoints and angles. Notably congested, the dataset exhibits head counts ranging from 49 to 12,865.

Then they compared on them their method and other known from literature:

* On the ShanghaiTech Dataset CL-DCNN got the best MAE on part A (52.6)
* On the Mall Dataset CL-DCNN got the best MAE (1.55) and RMSE (2.01).
* On the UCF\_CC\_50 Dataset CL-DCNN got best MAE (181.8)
* On the UCF-QNRF Dataset CL-DCNN got best MAE (96.4) and RMSE (168.7)

The authors made also some comparison with different number of DCMs. For part A of ShanghaiTech Dataset the best is one, and for part B the best is three. They also checked the Backbones such as AlexNet, VGG16 and ResNet50. VGG16 got the best results on both parts of ShanghaiTech Dataset.

This study presents CL-DCNN, a dilated CNN enhanced with a contextual module to capture both global and local features. It outperforms current methods on five public datasets while keeping computational costs reasonable. However, some limitations remain, and future work should focus on combining the model with lightweight CNNs for use in practical settings like UAVs.