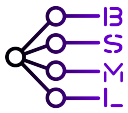


Crowd Counting through Density Map Estimation

Flavio Caroli* Luca Colaci* Vittorio Rossi*

Bocconi Students for Machine Learning, Bocconi University, Milan, Italy
{first.author second.author third.author}@studbocconi.it

November 4, 2025

Abstract

Crowd counting via density map estimation is highly sensitive to a model's receptive field, which determines how much spatial context informs each prediction. We present a controlled study on ShanghaiTech Parts A (dense) and B (sparse) that isolates the effect of receptive field by varying depth in UNet-style encoder–decoder architectures with pretrained VGG19 and ResNet50 backbones. Ground-truth density maps are generated with geometry-adaptive Gaussians, and we evaluate both count- and pixel-level errors (MAE/RMSE). Our modified UNet outputs half-resolution density maps and uses skip connections after max pooling to focus the analysis on receptive-field behavior. Results show a clear data–architecture match: on dense Part A, VGG-D4 attains the best count accuracy, benefiting from strong local feature extraction; on sparse Part B, ResNet-D4 performs best, leveraging a larger effective receptive field to suppress false positives in empty regions. Deeper variants generally improve density fidelity across both families. We also report that naive patch-based augmentation increases sample count but harms validation generalization due to distribution shift.

1 Introduction

Crowd counting plays a pivotal role in public safety, urban planning, and autonomous systems by estimating the number of individuals in a scene. A widely adopted formulation of this task is density map estimation, where a model predicts a spatial distribution of crowd presence rather than discrete counts [4]. However, performance in such tasks is often highly sensitive to the receptive field of the network, which governs how much spatial context a neuron can leverage when interpreting complex visual scenes [2]

In this work, we investigate how varying the depth of models (then the receptive field) impacts the quality of density maps and, consequently, the final crowd count accuracy. We use the ShanghaiTech dataset and adapt it to a density estimation setting using established Gaussian kernel techniques. We adopt an encoder-decoder structure inspired from UNet [3] on which we mount custom encoding blocks of two different categories: one with encoding blocks taken by ResNet50 and one with encoding blocks taken from VGG19.

Our aim is to empirically characterize how receptive field configurations influence density map fidelity and crowd counting performance, offering insights into architectural decisions for vision tasks involving spatially distributed labels. Our analysis is tied to the results reported on the ShanghaiTech dataset, particularly Parts A and B, as documented in recent benchmarks [1].

2 Problem Analysis

The crowd counting task, particularly in dense and unstructured scenes, poses several challenges: large variations in scale, severe occlusions, and uneven crowd distributions make direct regression approaches prone to errors. Density map estimation offers a spatially-aware solution, but its quality is closely tied to the network's ability to capture local features and global context elements governed primarily by the receptive field.

We hypothesize that models with larger recep-

*Equal contribution, the ordering is alphabetical.

tive fields will better handle sparse scenes and provide smoother, more coherent density maps due to their capacity to integrate broader context [6]. In contrast, smaller receptive fields might better preserve local density peaks in congested areas but could struggle with scattered or large-scale crowds. By varying receptive field sizes—through changes in architectural depth—we aim to explore these trade-offs systematically.

3 Methodology

Our approach investigates how receptive field size affects the performance of crowd density map estimation and counting. We construct a controlled experimental setup using the ShanghaiTech dataset and design multiple network variants differing in their receptive field configurations.

3.1 Dataset preparation & tentative augmentation

We utilise the ShanghaiTech dataset, which includes two parts: Part A, composed of images from highly congested internet scenes, and Part B, with relatively sparser images from Shanghai streets. We then follow the standard preprocessing pipeline by converting each annotated head point into a density map via geometry-adaptive Gaussian kernels, as proposed in [4]. This adaptation enables spatially aware supervision rather than scalar count regression, which has been shown to improve robustness in dense crowd scenarios [5].

We also tried augmenting the dataset by taking image patches (possibly overlapping to grant of fixed size, and producing ground truth maps accordingly. This allows to increase the size of our dataset $\approx 6\times$ the original image count. However, the model trained on the augmented dataset had severe issues generalizing to the validation images, as generated densities where not representative of the validation dataset.

3.2 ResNet50 & VGG19

Modules `ResNetSkip1` and `VGGSkip1` modules adapts a pretrained encoders to produce a density map at half the input resolution. These blocks are made by extracting the pretrained models' weights and adapting them to our desired input/output sizes. ResNet50 comprises approximately 25.6 million parameters and features a receptive field of 483 pixels, enabling it to capture broad contextual information efficiently. In contrast, VGG19 contains about 144 million parameters with a receptive field of 212 pixels.

3.3 UNet Structure

The pretrained blocks are composed in an encoder-decoder fashion, with MaxPooling layers reducing the dimensionality of the features. The encoder, on the other hand, works through a series of double 3×3 convolutions preceded by a Bilinear Upsamling layer. We instatiate a series of skip connections between encoder and decoder blocks at the same level.
Differently from classical UNet architecture, we made a couple of design choices due to both trial and error on a validation subset and due to the specific goal of our study. First, we decided not to instatiate a full decoding part of UNet, therefore ending up with density maps at half resolution. Second, all skip connections starting from the encoder are given after the MaxPooling layer, differently from the original architecture that takes them before MaxPooling, just after ReLU activation.

3.4 Receptive Field in Convolutional Neural Networks

In Convolutional Neural Networks (CNNs), of an output neuron refers to the region of the input image that influences the value of that neuron. It quantifies how much of the input is visible to a particular output activation. For a given layer l, the receptive field R_l can be recursively computed from previous layers as follows:

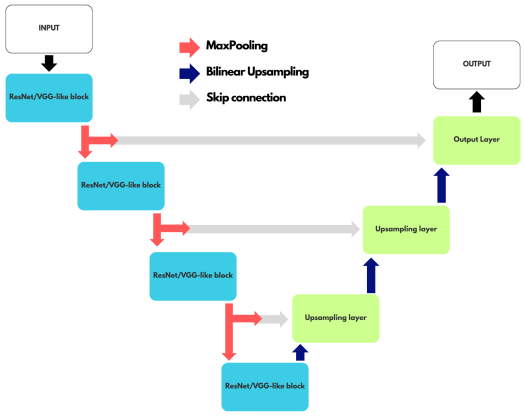
$$R_l = R_{l-1} + (k_l - 1) \cdot j_{l-1}$$

where k_l is the kernel size at layer l, and j_{l-1} is the jump (i.e., the effective stride) from the input to layer $l-1$. The jump is updated at each layer as:

$$j_l = j_{l-1} \cdot s_l$$

where s_l is the stride at layer l. The initial receptive field R_0 is 1, and $j_0 = 1$.

Filter size increases the receptive field linearly: larger kernels allow each neuron to aggregate information from a wider area of the input. We utilized two backbone modes with different initial receptive fields and augmented their depth to further investigate how these differences affect performance in pixel-wise density map estimation and counting.



4 Experiments and Metrics

We define several model variants depend on depth in the encoder and decoder blocks which affect significantly the receptive fields of the final layers of the modls (cf. [2]).

We report both *count* errors and *pixelwise* errors, using:

MAE & RMSE: *Count MAE* is computed by summing each density map and taking the mean absolute difference between predicted and ground-truth counts.

Count RMSE is the root of the mean squared difference between total predicted and ground-truth counts.

Pixel MAE is the average absolute per-pixel error between predicted and ground-truth density maps.

Pixel RMSE is the root of the mean squared per-pixel error.

Baselines are produced to give a sense of the goodness of the models: *Mean* is the goodness of predicting a total count which is the average across the dataset, spreaded across the images' pixels; *Zeros* is the result of predicting an empty density map.

5 Results and Discussion

The performance differences between VGG and ResNet on Parts A and B can be attributed to the distinct characteristics of each dataset subset and the architectural properties of the networks:

Part A - Dense Crowd Scenarios: Part A contains images with high crowd density where people are closely packed and often overlapping. In these scenarios, VGG's architecture proves more effective because: Dense crowds requires more precise feature extraction at multiple scales, also VGG's more direct feature extraction pathway and of course its higher number of parameters is better suited for capturing the local patterns and textures that distinguish individual people in crowded scenes. Finally, the high density means that local contextual information is more important than long-range spatial relationships.

Part B - Sparse Crowd Scenarios: Part B contains images with lower crowd density and more empty zones, where ResNet's superior performance can be explained by: **Larger Receptive Field**: ResNet's residual connections and deeper architecture provide a significantly larger effective receptive field, allowing the network to capture broader spatial context. **Empty Zone Handling**: The presence of large empty areas in Part B images requires the model to understand global scene context to avoid false positives in empty regions.

The depth analysis shows that deeper networks (D4) generally perform better, with VGG-D4 excelling on Part A and ResNet-D4 on Part B, suggesting that increased model capacity benefits both architectures when matched to appropriate scenarios.

These results highlight the importance of matching network architecture to dataset characteristics: VGG excels in dense scenarios requiring fine-grained local analysis, while ResNet's broader receptive field and residual learning are advantageous for sparse scenarios requiring global context understanding.

6 Conclusion

In this work, we investigated the impact of receptive field size on crowd density map estimation by systematically varying the depth of VGG19 and ResNet50-based encoder-decoder architectures. Our experiments on the ShanghaiTech dataset reveal that architectural choice should be tailored to crowd density characteristics. The findings highlight the trade-off between local precision and global context aggregation and receptive field characteristics in density map estimation and counting.

References

[1] State-of-the-art crowd counting on shanghaitech a and b. https://paperswithcode.com/sota/crowd-counting-on-shanghaitech-a, 2025.

Table 1: Crowd counting performance by various models and baselines on ShanghaiTech Parts A and B. Best results are highlighted in **bold**.

Metric	**Baseline**		**VGG**			**ResNet**		
	Mean	Zeros	D2	D3	D4	D2	D3	D4
Part A								
MAE (count)	254.39	432.89	171.50	170.27	**109.30**	227.18	144.73	120.81
RMSE (count)	353.17	558.68	239.10	203.31	**150.41**	313.83	193.54	166.49
MAE (pixel)	1.0e−3	1.0e−3	8.71e−3	6.95e−3	5.99e−3	1.047e−2	**5.77**e−**3**	6.38e−3
RMSE (pixel)	4.0e−5	5.0e−5	1.7e−2	1.39e−2	1.22e−2	6.7e−5	**1.2**e−**2**	1.39e−2
Part B								
MAE (count)	71.95	123.56	123.10	21.85	37.01	56.19	37.98	**19.82**
RMSE (count)	88.48	155.87	147.25	26.25	42.76	65.84	42.85	**24.81**
MAE (pixel)	1.0e−3	1.1e−3	3.3e−3	1.8e−3	1.58e−3	2.1e−3	1.96e−3	**1.5**e−**3**
RMSE (pixel)	9.0e−6	1.1e−5	1.1e−2	5.3e−3	4.5e−3	5.9e−3	5.5e−3	**4.3**e−**3**

[2] Xiaoshuang Chen, Yun Zhao, Yu Qin, Fei Jiang, Mingyuan Tan, Hongtao Lu, and Xiansheng Hua. Panet: Perspective-aware network with dynamic receptive fields and self-distilling supervision for crowd counting. *arXiv preprint arXiv:2111.00406*, 2021.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[4] Jinjun Wan and Antoni B Chan. Adaptive density map generation for crowd counting. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[5] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.

[6] Yujun Zhang, Yifan Zhou, Yao Chen, Xinyu Gao, and Yuesheng Xu. Atrous convolutions spatial pyramid network for crowd counting and density estimation. *Neurocomputing*, 365:139–147, 2019.

A Results

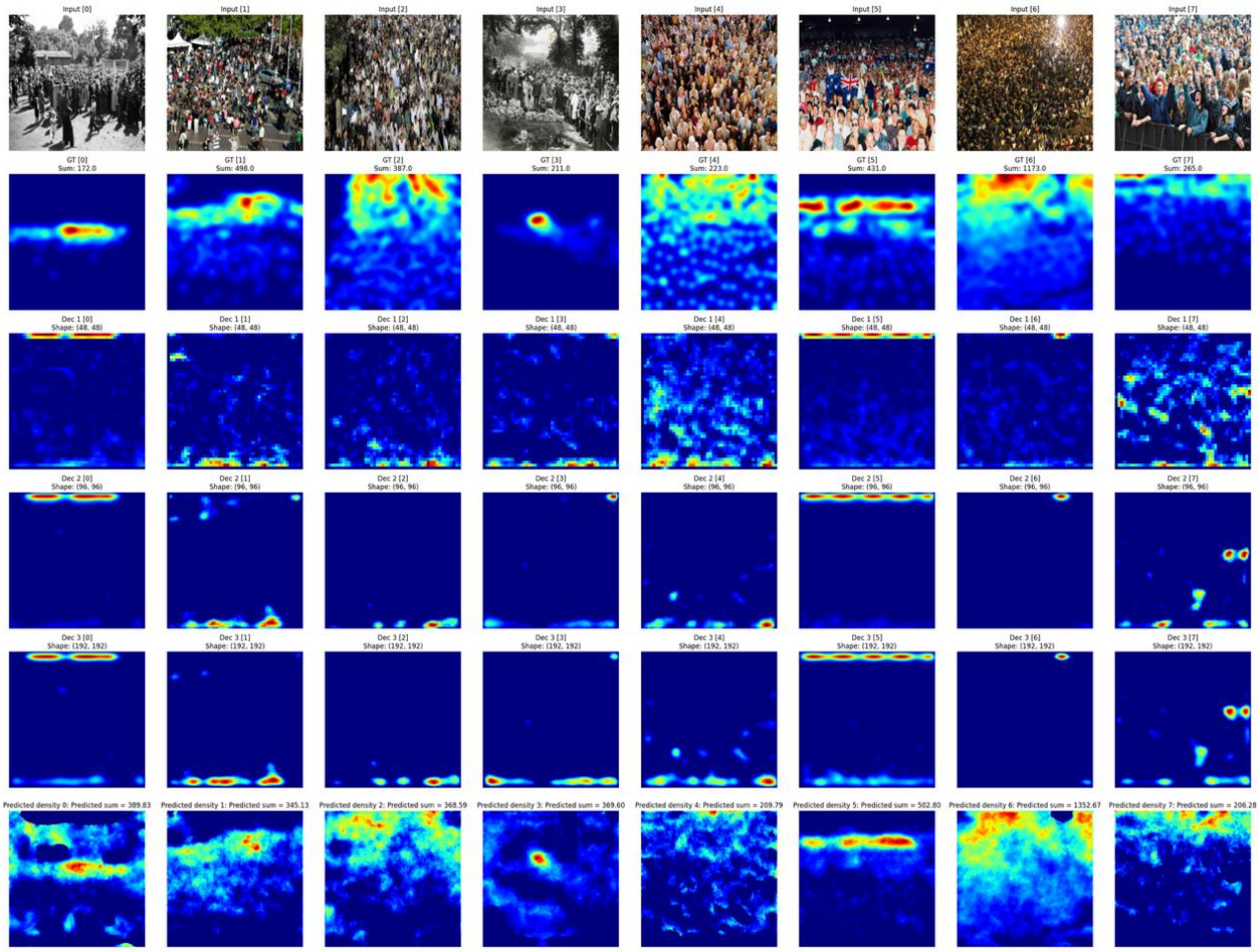


Figure 1: Density map visualization for learned filters VGG-D4 on ShanghaiTech Part A.

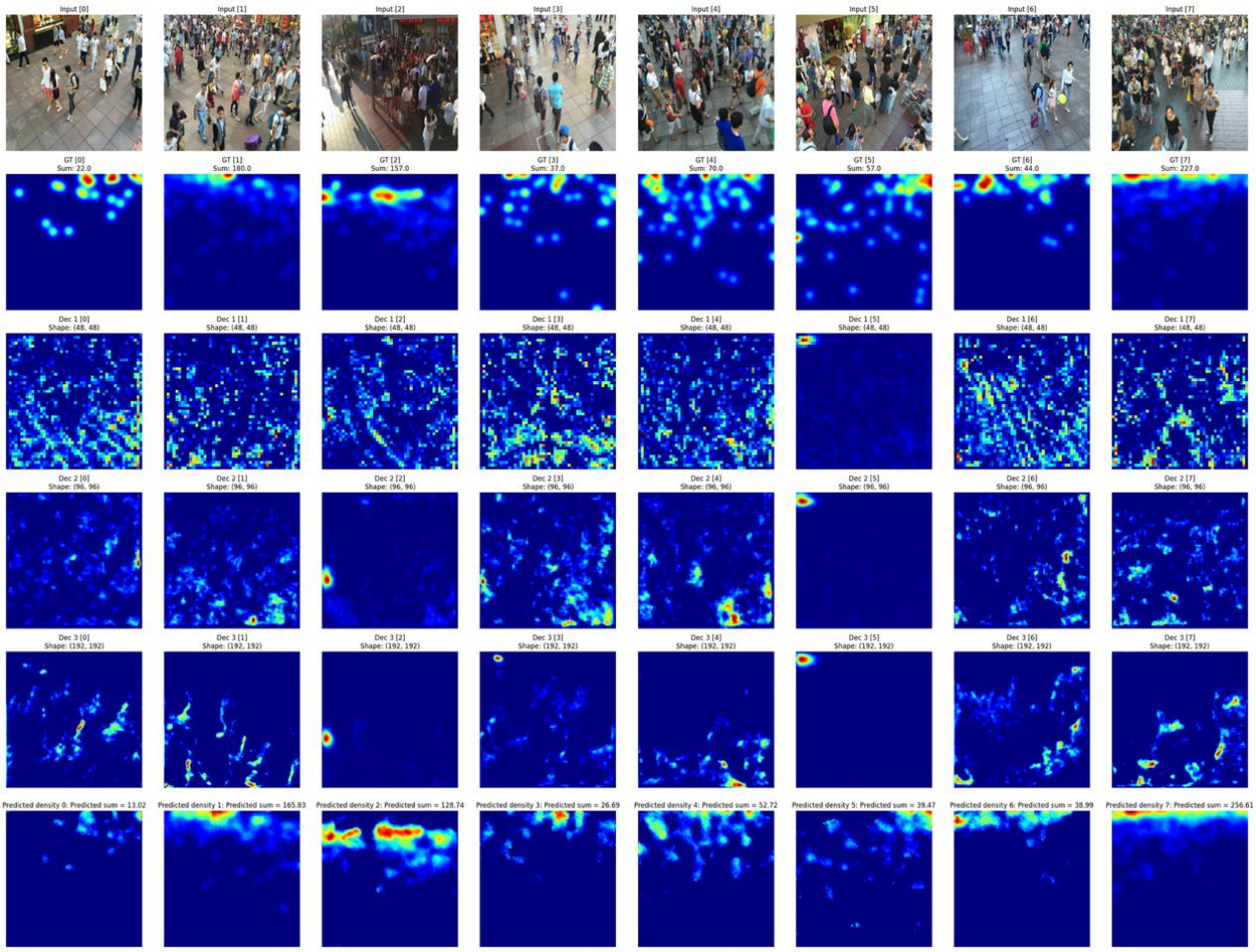


Figure 2: Density map visualization for learned filters in ResNet-D4 on ShanghaiTech Part B.

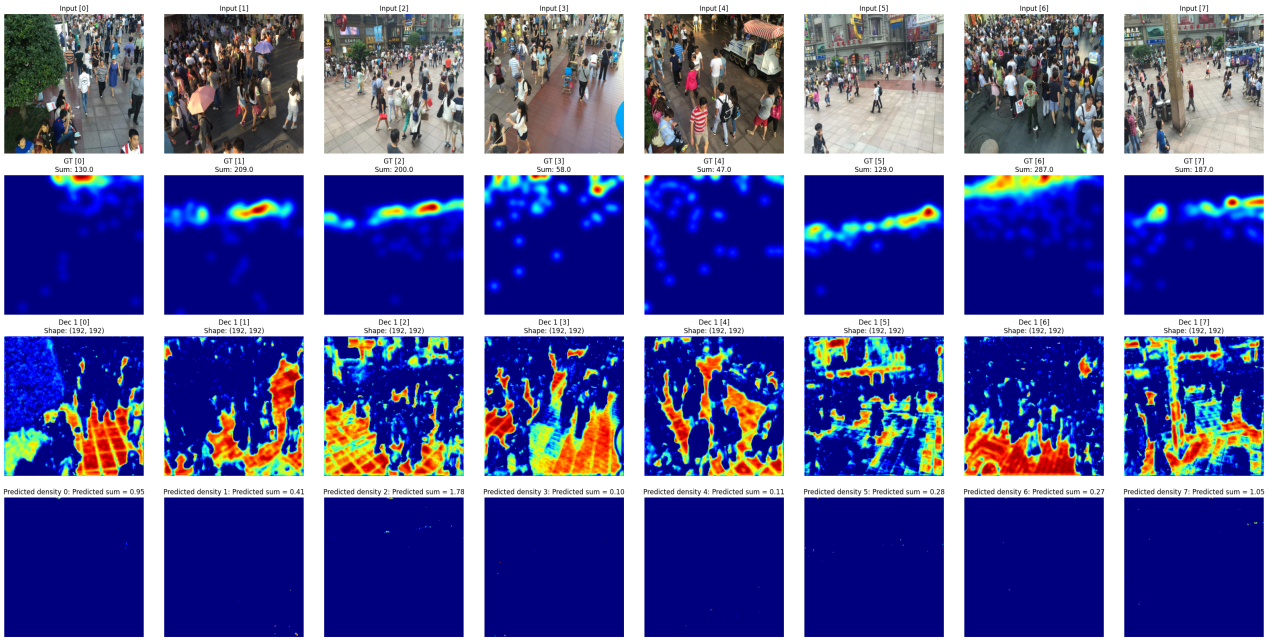


Figure 3: Density map visualization for learned filters in VGG-D4 and problems in last output on ShanghaiTech Part B.