

# Different ways to interpret Crowd Counting

Follador Alessandro

alessandro.follador@studenti.unipd.it

Franzin Francesco

francesco.franzin.1@studenti.unipd.it

Toffolo Nicol 

nicolo.toffolo.1@studenti.unipd.it

## Abstract

*This study evaluates various computer vision methodologies for crowd size estimation within a specific visual field. Utilizing footage from stationary surveillance cameras in a retail environment, the research explores the practical utility of these techniques. Key applications include enhancing public safety, analyzing visitor traffic for commercial A/B testing, optimizing architectural space management, and monitoring adherence to public health protocols.*

## 1. Introduction

Crowd counting is a specific task within the field of Computer Vision that involves automatically estimating the number of individuals in a digital image or video sequence. Unlike simple object detection, crowd counting must often deal with highly congested environments where individuals are partially obscured (occlusion), vary significantly in size due to perspective (scale variation), and appear in diverse lighting conditions. The main goal of crowd counting is to generate an accurate estimate of the total population in a given area.

In this study, we evaluate a spectrum of methodologies, ranging from foundational regression models to cutting-edge transformer-based architectures. The core aim is to investigate whether simpler architectures, when augmented with State-of-the-Art (SOTA) training strategies, can achieve performance levels comparable to more structurally complex models in scenarios where precise spatial labels are unavailable. To this end, we analyze four approaches.

We start from an adaptation of the Multi-column Convolutional Neural Network (MCNN) [1], to our specific task and dataset. Generally, these models are used to generate density maps by leveraging point-level annotations (typically the coordinates of each head) to handle extreme variations in perspective and crowd density. However, in our specific implementation, the MCNN is adapted to work

with a global count supervision, effectively transforming the spatial density estimation task into a distributed regression problem across the output grid. By doing so, we aim to evaluate whether the structural advantages of a multi-column design can still provide robust feature extraction.

Then we proceed with an approach that actually treats crowd counting as a regression task. The model built upon the ResNet-18 architecture [2], simplifies the problem by mapping high-level visual features directly to a scalar count.

The third approach is a SOTA-Enhanced ResNet-18, this model maintains the ResNet-18 backbone but incorporates State-of-the-Art (SOTA) training strategies, leveraging Transfer Learning.

The last model we consider is an adaptation of the PET Quadtree (Point-Query Transformer), that is the current State-of-the-Art, which moves away from traditional convolutions toward a Transformer-based approach[3].

## 2. Related Works

The field of crowd counting has evolved from traditional computer vision techniques to sophisticated deep learning architectures. This section outlines the typologies of models considered by us, that goes from global regression to spatial density estimation and, finally, to the transformer-based localization.

**Residual learning and global regression.** The emergence of Deep Convolutional Neural Networks (CNNs) revolutionized feature extraction. A pivotal milestone was the introduction of ResNet by He et al. [2]. By implementing residual blocks with shortcut connections, the authors addressed the vanishing gradient problem, allowing for the training of significantly deeper networks.

In the context of crowd counting, ResNet-18 can be adopted as a robust backbone for global regression tasks. Unlike classification, where the network identifies a category, regression-based approaches map high-level spatial features directly to a scalar value representing the total count.

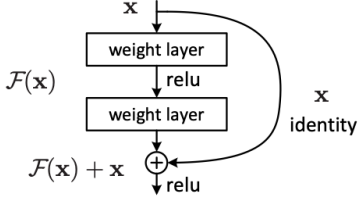


Figure 1. Residual learning: a building block.

**Density-based approaches.** To overcome the loss of spatial information inherent in global regression, researchers shifted toward density map estimation. The Multi-column Convolutional Neural Network (MCNN), proposed by Zhang et al. [1], addressed the challenge of scale variation. By utilizing three parallel columns with varying filter sizes, MCNN could simultaneously capture features of different head sizes.

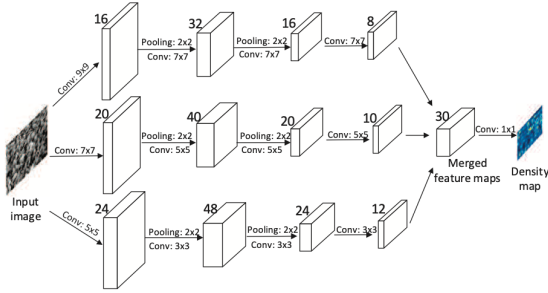


Figure 2. Structure of MCNN for density map estimation

This work popularized the use of geometry-adaptive gaussian kernels to transform point annotations into continuous density maps, which has since become a standard evaluation protocol in the field.

**Transformers and point-query localization.** The current frontier of crowd counting moves away from pixel-wise density estimation toward set-prediction and localization. The Point-Query Transformer (PET) with Quadtree refinement, introduced by Liu et al. [3], represents the current State-of-the-Art. By leveraging the self-attention mechanism of Transformers, PET captures global dependencies that CNNs might miss. The integration of a Quadtree structure allows for recursive spatial partitioning, concentrating computational resources on dense areas.

This approach effectively bridges the gap between counting and individual localization, providing superior accuracy on high-congestion datasets.

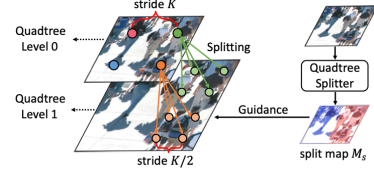


Figure 3. Illustration of point-query quadtree

### 3. Dataset

The Mall Dataset was introduced by Chen et al. [4] in 2012 and has since become a standard benchmark for crowd counting in low-to-medium density indoor environments. The data was captured using a publicly accessible surveillance camera installed in a shopping mall. It consists of 2,000 frames extracted from the original video sequence, so it provides a consistent background and stable lighting conditions, though it introduces challenges related to perspective and scene-specific occlusion. Each frame has a resolution of 640×480 pixels and there are over 60,000 individual head instances annotated across all of them. The number of people per frame varies from 13 to 53, with an average of 31 individuals.

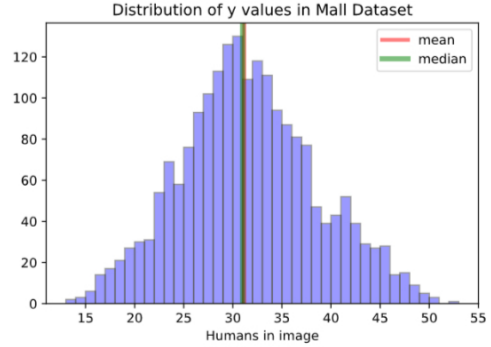


Figure 4. Mall dataset target value distribution

Due to the high-angle positioning of the camera, individuals at the top of the frame appear smaller than those in the foreground. This perspective effect provides an ideal benchmark for evaluating scale-aware models, such as MCNN [1].



Figure 5. A sample from the Mall dataset

**Data partitioning.** For the experimental phase, the Mall Dataset was partitioned into three distinct subsets. Out of the 2,000 total frames, we adopted a distribution strategy

that prioritizes a larger training set of 1,360 images (68 percent of the total), while maintaining representative samples for validation and final testing, consisting of 240 images (12 percent) and 400 images (20 percent) respectively.

## 4. Method

### 4.1. MCNN adaptation

Given that the dataset lacks point-level coordinates (x,y) for individual heads, we adapted the MCNN to work with a Uniform Ground Truth. The global count label for each image is distributed equally across a  $64 \times 64$  grid (representing  $1/4$  of the input resolution). Each pixel in this target map is assigned a value of  $\frac{Count}{(64 \times 64)}$ , followed by a light Gaussian smoothing ( $\sigma=1.0$ ) to facilitate gradient flow during training.

The MCNN was trained from scratch using random weight initialization. Under this configuration, the model acts as a distributed spatial regressor. Instead of learning to localize individuals, it learns to correlate specific visual patterns, such as the frequency of edges and textures associated with a crowd, to a local density value. Essentially, the network estimates an average number of people per pixel based on the visual complexity of the scene.

The final global count is then approximated by summing these local predictions across the entire output map.

### 4.2. ResNet-18

To better understand the impact of prior knowledge on crowd counting accuracy, we implemented and compared two versions of the ResNet-18. Both models share the same fundamental architecture but differ in their initialization and training strategy.

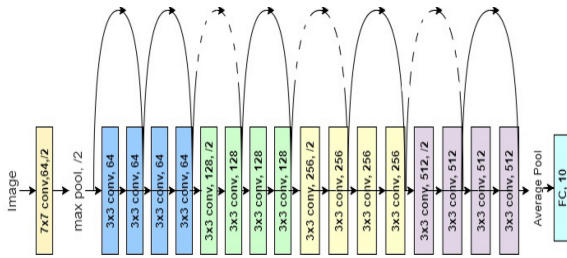


Figure 6. ResNet-18 architecture

Initially, we tested a ResNet-18 initialized with random weights, with the aim to evaluate the network’s ability to learn crowd-specific features. The final classification layer (fc) is replaced with an Identity block to extract a high-level feature vector. A custom Regression Head is attached, consisting of a Linear layer (128 units), ReLU activation, Dropout for regularization, and a final Linear layer that outputs a single numerical value. One of the most important technical choices in this code is the use of the Log-Space

Transformation. Instead of predicting the raw count, the model predicts  $\log(1 + count)$ , the reason is because the Crowd data is often skewed a few images might have hundreds of people while most have few. Using the log function compresses the range of the target values, making the training more stable and helping the MSE (Mean Squared Error) loss converge more effectively.

Then we transitioned to a Pre-trained ResNet-18, which utilizes weights previously optimized on the ImageNet dataset. By starting from these weights, the model focuses its training on “refining” its knowledge for the Mall dataset rather than learning basic visual structures from scratch.

### 4.3. SOTA-Enhanced ResNet-18

Unlike the baseline approach, this model leverages Transfer Learning by employing a ResNet-18 backbone pre-trained on the ImageNet-1K dataset. By utilizing pre-learned visual representations, ranging from low-level edges and textures to high-level geometric shapes, the model achieves significantly faster convergence and a superior understanding of spatial patterns compared to training from scratch. This hierarchical feature extraction allows the regressor to focus on density-specific patterns, improving the overall counting accuracy even with limited domain-specific data.

To mitigate the inherent long-tail distribution typical of crowd counting datasets, where low-count images significantly outnumbered high-density scenes, the training pipeline implements a WeightedRandomSampler. The dataset is partitioned into frequency-based bins according to crowd density; rare samples, such as images containing dense crowds, are assigned higher sampling weights. This technique ensures a balanced exposure during training, forcing the model to learn high-density features as effectively as low-density ones and preventing a predictive bias toward lower numerical values.

### 4.4. PET Quadtree

The implemented PET (Point-and-point Extraction Transformer) architecture represents an adaptation of the original work [3], customized to combine the effectiveness of Transformers with the constraints of the Mall Dataset and available computational resources.

This architecture shifts the crowd counting paradigm from global regression to a point-set prediction task. Instead of estimating a single density value, the model attempts to localize individuals as discrete points in space, providing both a total count and spatial distribution.

**Architecture.** The model utilizes a ResNet-18 backbone as its primary feature extractor. By leveraging the first three layers of the ResNet, the network generates a high-dimensional feature map that preserves the spatial geome-

try of the input image. This map is then projected into a “Transformer dimension”, allowing the subsequent stages of the model to interpret visual patterns (such as heads and shoulders) as potential point candidates.

A key innovation in this implementation is the integration of a Quadtree Splitter. Recognizing that crowd density is rarely uniform, the model dynamically divides the image into a grid of cells. A dedicated “Split Head” analyzes each cell and if a region is identified as high-density, the quadtree recursively subdivides it into four smaller quadrants. This allows the model to allocate more “queries” (computational focus) to crowded areas while remaining efficient in sparse regions, effectively handling the large variations in scale and perspective found in the Mall dataset.

For each cell generated by the quadtree, the model extracts a local feature token. These tokens are fed into a Transformer Decoder, which performs a global reasoning task by comparing each token against the entire image’s feature memory. Through this mechanism, the model refines the probability and the exact (x,y) offset of each point. The final prediction head outputs a set of logits for classification (detecting the presence of a person) and regression offsets for precise positioning.

A distinctive aspect of this adaptation is its approach to training under weak supervision. Since the dataset lacks precise ground-truth coordinates, the model generates uniform random points based on the total count for each frame. To bridge the gap between these synthetic points and the model’s predictions, the Hungarian Matching algorithm [5] is employed. By solving a bipartite matching problem, the model learns to assign each prediction to a unique ground-truth point.

A second implementation optimizes the Point-and-point Extraction Transformer by replacing manual patch extraction with Bilinear Grid Sampling. By mapping quadtree centers directly into the feature space. The model achieves sub-pixel precision and higher computational efficiency during the feature-to-query transformation. The training process utilizes a Hybrid Supervision strategy:

- Hungarian Matching: Ensures one-to-one correspondence between predictions and points, preventing multiple detections for the same individual.
- Global Count Anchor: A global Smooth L1 loss is applied to the sum of predicted probabilities. This enforces global consistency, ensuring the total “activation energy” of the network aligns with the actual crowd count.

During inference, the model adopts Probabilistic Counting, summing the sigmoid scores of all queries. This approach provides a smoother, more robust estimate than hard thresholding, effectively lowering the Mean Absolute Error

(MAE) by better accounting for visual uncertainty in dense areas.

#### 4.5. Performance evaluation

For the entire project, we adopted the Mean Absolute Error (MAE) as the evaluation metric. The main reason for this choice is that, in the context of crowd counting, MAE directly represents the average deviation of the predicted count from the ground truth. The MAE is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

where  $y_i$  denotes the ground truth count for the i-th image,  $\hat{y}_i$  is the corresponding prediction produced by the model, and  $N$  is the total number of samples. For instance, an MAE value of 5 indicates that, on average, the model’s prediction deviates from the true count by five individuals. This makes MAE highly intuitive and easily interpretable. Furthermore, MAE treats all errors linearly.

In crowd counting scenarios, some images may produce extremely large prediction errors (outliers); compared to Mean Squared Error (MSE), MAE is less sensitive to such isolated large errors and therefore provides a more stable assessment of average model performance.

### 5. Experiments

In this section we present the results obtained for each model.

Architecture	MAE
MCNN adapted architecture	5.5435
MCNN adapted architecture (Count Regressor)	31.1375
ResNet-18 Regressor (no pre-train)	2.1747
ResNet-18 Regressor (pre-train)	1.7671
ResNet-18 Regressor (pre-train, TTA, Weighted Sampling, SmoothL1Loss, CosineWarmup)	2.3097
PET Quadtree (no pre-train)	10.5425
PET Quadtree (pre-train, with parameter changes)	2.0908

Table 1. Comparison of Models Performances (MAE)

The results are reported in Table 1, from which significant differences among the evaluated models can be clearly observed.

#### 5.1. MCNN based approaches

First, both MCNN-based approaches exhibit relatively high MAE values, even after adapting the architecture to our dataset. In particular, the MCNN Count Regressor performs poorly, suggesting that directly regressing the crowd count without explicitly modeling spatial density is not well suited for this task in our setting.

## 5.2. ResNet-18 based approaches

In contrast, the ResNet-based regressors achieve substantially better performance. Notably, the introduction of pre-training leads to a clear improvement, reducing the MAE from 2.17 to 1.77. This highlights the importance of transfer learning, as the pre-trained backbone provides more robust and discriminative visual features. Interestingly, the most complex ResNet-18 configuration, which includes test-time augmentation, weighted sampling, SmoothL1 loss, and cosine warmup scheduling, does not outperform the simpler pre-trained version. This suggests that, given the limited variability and size of the dataset, additional training strategies may introduce unnecessary complexity without yielding further benefits causing some over-fitting problem with an high variance.

## 5.3. PET Quadtree

For the PET Quadtree approach, the model trained without pre-training shows a relatively high MAE. This method operates by identifying and classifying individual points corresponding to human heads and then summing the associated probabilities to obtain the final count. When pre-training and parameter tuning are introduced, the MAE decreases significantly, confirming once again the effectiveness of pre-trained representations.

## 6. Conclusions

In conclusion the best model we found is the ResNet-18 utilizing pre-training.

Focusing the performance of this model, we further analyzed the six worst predictions, that can be seen in Figure 7. In all these cases, the crowd distribution within the image it is highly close-up with large empty areas. The model appears to struggle when crowd density varies significantly within the same scene.



Figure 7. Six worst predictions

In particular, the images with highest MAE (with errors

of 6 and 7) contain tightly packed groups of individuals concentrated around central points of interest. This suggests that the model has difficulty distinguishing individuals when they are partially occluded or located very close to each other. Nevertheless, the overall MAE of approximately 1.8 on scenes containing between 30 and 42 people indicates a very low relative error. Moreover, the predicted counts vary substantially (ranging from 24 to 49), closely following the true variations in the ground truth (from 30 to 42). If the model were merely learning the dataset mean, the predictions would instead be much more uniform and clustered around a single average value.

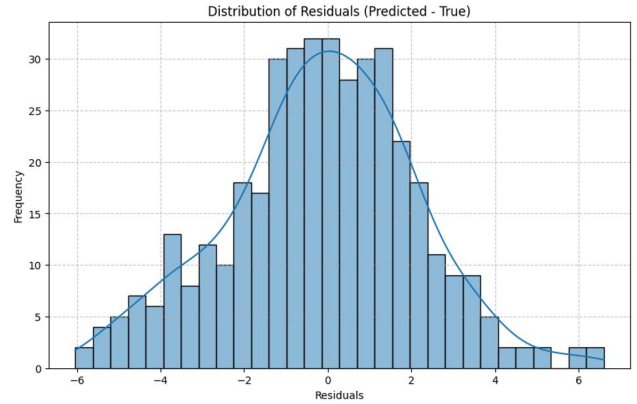


Figure 8. A sample from the Mall dataset

These observations confirm that the model effectively learns meaningful visual cues from the images rather than relying on a simple average-based prediction strategy. An MAE below 2 in a crowd counting scenario with an average of approximately 30 people per image corresponds to a relative error of less than 7 percent, which is highly competitive given the visual complexity of the scene and the presence of occlusions.

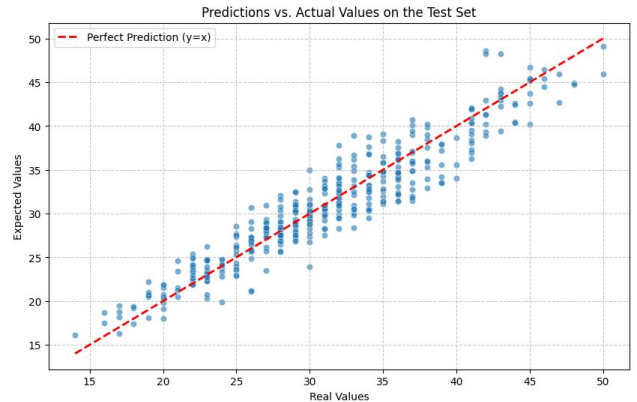


Figure 9. A sample from the Mall dataset

The comparative analysis reveals that for crowd counting tasks characterized by mid-to-low density datasets, a

backbone such as ResNet-18 provides sufficient representative power to achieve high accuracy. The experimental results demonstrate that adapting the MCNN (Multi-Column Convolutional Neural Network) structure for direct regression does not yield satisfactory results, as evidenced by its significantly higher error rates. While MCNN was originally designed to handle head size variations through multiple columns, its shallower architecture struggles to compete with the global feature extraction capabilities of more modern residual networks in a regression context.

In contrast, the PET Quadtree architecture represents a more complex and computationally intensive approach. While it shows improvement when pre-trained, its sophisticated quadtree and transformer-based mechanisms are better suited for high-density datasets or tasks requiring precise localization, such as individual head detection and point-level density estimation in extremely crowded scenarios.

Finally, the study highlights the critical influence of Transfer Learning; leveraging weights pre-trained on ImageNet allows the models to bypass the difficulties of random initialization, leading to faster convergence and a deeper understanding of visual spatial patterns that are essential for accurate crowd estimation.

## References

- [1] Yingying Zhang, Desheng Zhou, Si Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. pages 589–597, 2016.
- [2] Kaiming He, Xiangyu Zhang, Ren Shaoqing, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.
- [3] Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongming Liu. Point-query quadtree for crowd counting, localization, and more. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [4] Ke Chen, Chen Change Loy, Shaogang Gong, and Tao Xiang. Feature mining for localised crowd counting. In *European Conference on Computer Vision (ECCV)*, pages 1–11. Springer, 2012.
- [5] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.