Solving Scale Variation in Crowd Density Estimation using CSRNet and CBAM

Mahirah Sofea [redacted] Julia Irsalina [redacted] Nur Mushira [redacted]

1. Introduction

The Congested Scene Recognition Network, CSRNet is a deep learning model that performs well in crowd density estimation by using dilated convolutions to capture patterns at different scales. However, it struggles when crowd size varies greatly within an image. To address this, we propose a global Gaussian technique during data preparation to help the model better supervise and capture blur in the front region. On the model side, we integrate a Convolutional Block Attention Module (CBAM) into CSRNet to help it focus on more important areas in both channel and spatial dimensions.

2. Problem Identification

While CSRNet handles varying crowd scales with dilated convolutions, it still struggles with scale variation, especially when people in the front appear large and those in the back are small. Since annotations only mark head centers and not head size, CSRNet tends to highlight the background while underestimating the dense front region. Existing geometry-adaptive methods like KD-Tree only consider local density (dense vs. sparse), not the spatial layout (front vs. back). This leads to poor supervision, causing the model to miss critical front-area details and produce less accurate density maps.

3. Technical Soundness

3.1 Baseline Model: CSRNet

We use CSRNet as our baseline because it performs well in crowd density estimation, as shown in the original paper by Li et al. [1]. The model is organized into two main components: the front-end and the back-end. The front-end is responsible for extracting important features that consist of the first 10 convolutional layers of VGG-16 and the output feature maps at one-eighth the resolution of the original input image.

The back-end generates high-quality density maps and accurate counts using dilated convolutions, which expand the receptive field without reducing resolution throughout the network to preserve spatial information. This component remains unchanged, as CSRNet [1] has demonstrated outstanding performance with the structure compared to more complex alternatives. However, CSRNet lacks attention mechanisms and struggles with scale variation, which we aim to improve.

3.2 Proposed Method 1: Global Gaussian

Inspired by the geometry-adaptive kernel from Zhang et al. [1]., we propose using a global σ instead of assigning a different σ per head. We first compute the average distance between each head and its k-nearest neighbors using a KD-Tree, then take the mean of these distances to get a single global σ:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x), \quad \text{with} \quad \sigma = \beta \cdot \frac{1}{N} \sum_{i=1}^{N} \bar{d}_i$$

Our hypothesis is that by applying the same blur (σ) to all heads, we give equal importance to both *dense vs. sparse* and *near vs. far* regions. This allows the model to be supervised more consistently across the image, improving overall density estimation quality.

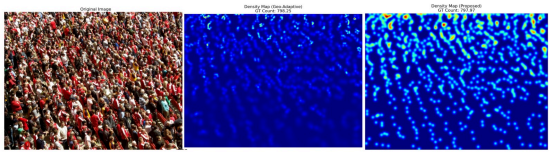

Figure 1: From left to right: original image, density map using geometry-adaptive Gaussian, and density map using proposed global Gaussian.

With the global Gaussian, the generated ground truth density map now highlights the front crowd, which was not emphasized in the geometry-adaptive approach. As a result, CSRNet is better able to recognize and learn the presence of front-region crowds during training.

3.3 Proposed Method 2: CBAM

As demonstrated in the research paper by Woo et al. [2] from Korea Advanced Institute of Science and Technology (KAIST), CBAM is a lightweight and general attention module that can seamlessly integrate into any convolutional neural network (CNN) architecture with negligible computational overhead. It is also fully end-to-end trainable alongside the base CNN.

This attention mechanism operates sequentially along two separate dimensions: channel and spatial. The resulting attention maps are multiplied by the input feature map to perform adaptive feature refinement. This enhancement guides the network on what to and where to focus so it

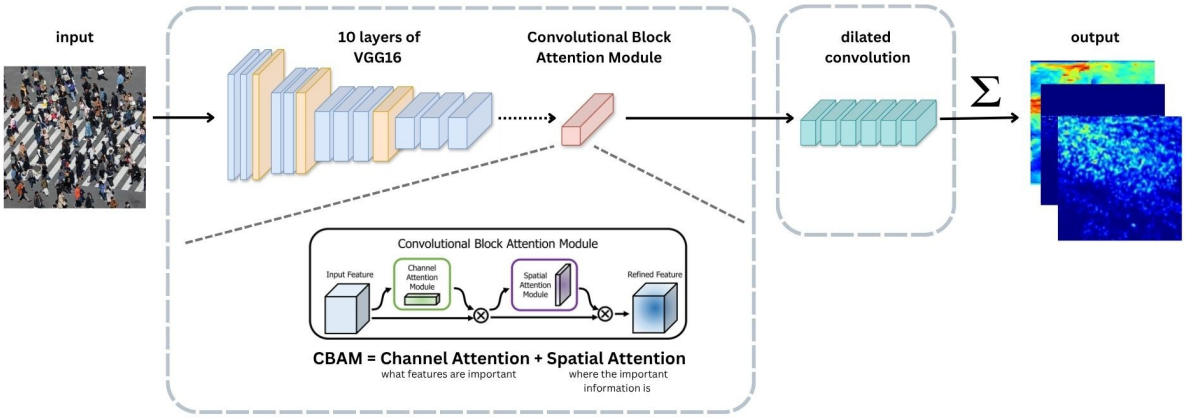

Figure 2: The proposed architecture of integrated Convolutional Block Attention Module (CBAM) adapted from Woo et al [2], and the original CSRNet architecture.

can improve the representation of important features while suppressing irrelevant information.

Our proposed model, as in Figure 2,implements CBAM to address the challenge of scale variation in crowd scenes. After extensive research and discussion, integrating the attention mechanism between the front-end and back-end would be most effective. The attention module enables the network to selectively focus more on informative regions (e.g., occluded heads) while ignoring irrelevant background elements (e.g., trees, empty areas). This enhancement of feature representations before the dilated convolutions ultimately leads to more accurate density map generation of crowd estimation.

Therefore, our hypothesis is that the generation of the density maps of crowd estimation could be improved and help to address the issue of scale variation as shown in figure below.

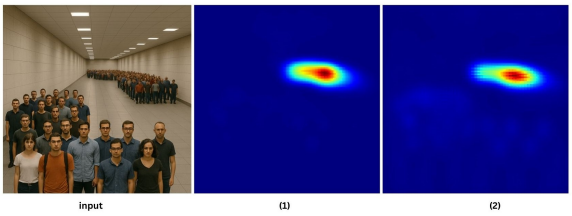

Figure 3: Generated density maps before (1) and after (2) the implementation of CBAM in the model

3.4 Loss Function

CSRNet originally uses Mean Squared Error (MSE) loss for density map regression. However, MSE alone may produce blurry outputs and ignores local structural details [5]. Following SANet [4], we apply a **Hybrid Loss** by adding the Structural Similarity Index (SSIM) to capture spatial structure.

Hybrid Loss. We combine MSE with SSIM to enforce both pixel accuracy and local structural consistency in density maps.

$$\mathcal{L}_{\text{hybrid}} = \text{MSE}(\hat{D}, D) + \lambda(1 - \text{SSIM}(\hat{D}, D))$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(p_i - g_i)^2$$

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Count Loss. To correct global count errors, we introduce Count Loss based on Mean Absolute Error:

$$\mathcal{L}_{\text{count}} = \frac{1}{N}\sum_{i=1}^{N}\left|C_{\hat{D}}^{(i)} - C_{D}^{(i)}\right|$$

Total Loss. Our final loss function combines both terms:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{hybrid}} + \beta \cdot \mathcal{L}_{\text{count}}$$

Justification. Table 1 shows that adding Count Loss to Hybrid Loss significantly lowers MAE and RMSE, and improves validation accuracy by combining structural and global count supervision.

Table 1: Effectiveness of Hybrid Loss + Count Loss

Loss Type	MAE	RMSE	Val Acc.
Hybrid Loss Only	434.30	647.87	15.80%
Hybrid + Count Loss	**103.76**	**181.67**	**79.71%**

4. Experimental Methodology

4.1 Datasets

The ShanghaiTech dataset contains 1,198 annotated images, divided into two parts: Part A, which includes 482 images of highly congested scenes, and Part B, which consists of 716 images of relatively sparse crowd scenes captured in street environments. However, in this project, we

use a subset of 250 images from Part A, split in a 1:4 ratio between testing and training, as we focus on solving challenges in high-density crowd scenarios.

4.2 Training Procedures

We aim to compare model performance using the same dilation rates but varying kernel sizes, applying both our proposed global Gaussian σ and the geometry-adaptive kernel from CSRNet [1] to both the baseline and proposed model. The Adam optimizer is employed with a batch size of 1, a learning rate of 1e-5, and a total of 30 training epochs. In each iteration, gradients are computed for the loss functions, and model parameters are updated accordingly. We evaluate the generated density maps by comparing the baseline and proposed models using the evaluation metrics described in Section 4.3. The complete test results are presented in Table 2.

4.3 Evaluation Metrics

We evaluate both count accuracy and density map quality using four metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR).

- **MAE** measures average error in predicted counts:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_{\hat{D}}^{(i)} - C_{D}^{(i)}\right|$$

- **RMSE** penalizes larger errors to capture variability:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_{\hat{D}}^{(i)} - C_{D}^{(i)}\right)^2}$$

- **SSIM** evaluates local structural similarity between predicted and ground truth density maps, reflecting how realistic the spatial layout appears.

- **PSNR** measures the peak signal quality of the predicted density map:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$$

  where MAX is the maximum possible pixel value of the image.

Together, these metrics assess both how accurate the model's predicted counts are and how visually realistic its density maps appear.

5. Result Analysis

5.1 Ablation Discussion

Table 2 shows results for four CSRNet variants trained with the same loss. Models differ only in the use of CBAM and Gaussian type.

Adaptive Gaussian yields the lowest MAE, while CBAM with global Gaussian gives the best RMSE, SSIM, and PSNR. This indicates a trade-off between count accuracy and density map quality.

Qualitative outputs show CBAM-based models focus more on background regions. All variants converge stably during training.

Table 2: Ablation results of Gaussian types and CBAM.

Backbone	Gaussian	MAE↓	RMSE↓	SSIM↑	PSNR↑
CSRNet	Adaptive	**93.657**	161.241	0.391	20.493
CSRNet	Global	114.920	187.923	0.454	23.716
CSRNetCBAM	Adaptive	103.755	181.668	0.350	20.416
CSRNetCBAM	Global	106.150	**159.459**	**0.576**	**26.238**

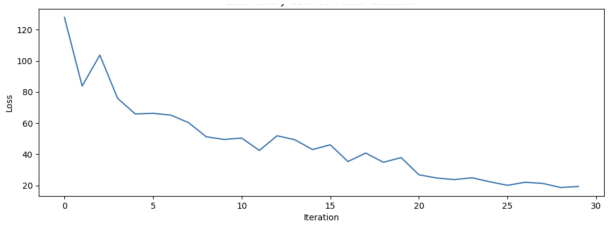

Figure 4: Loss history plot for CSRNet with Global Gaussian and CBAM.

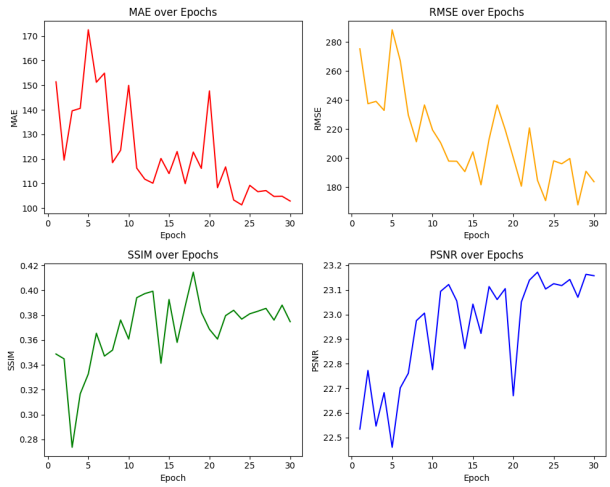

Figure 5: Evaluation metrics (e.g., MAE, MSE) for CSRNet with Global Gaussian and CBAM.

5.2 Scale Variation Analysis

We qualitatively assess scale variation using an image with large individuals in the front and smaller ones in the back. Figure 7 compares predictions from four models: CSRNet with adaptive Gaussian (baseline), CSRNet + Global, CBAM + Adaptive, and CBAM + Global tested on test image Figure 6

The baseline misses many front-region individuals. CBAM + Adaptive detects more people but often highlights full bodies. CBAM + Global gives clearer, head-focused predictions and performs well in both regions,

suggesting better spatial focus and improved handling of scale differences.

Figure 7 shows that our proposed model with CBAM and global Gaussian detects people in both the front and back, unlike the baseline which mostly detects the back. This shows improved robustness to scale variation.

Figure 6: Test image showing scale variation: large individuals in front, small in back.

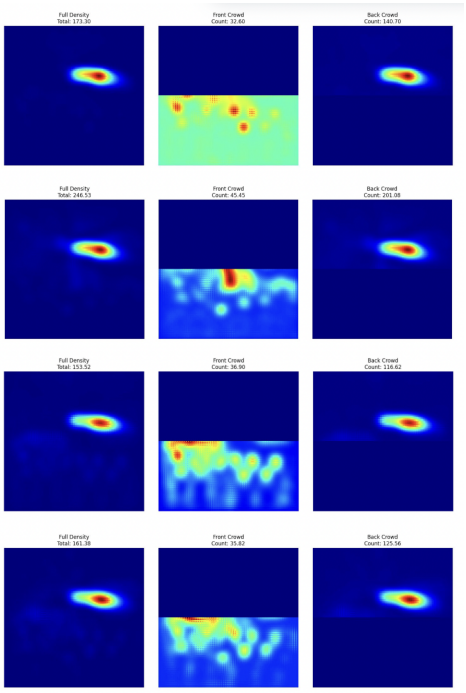

Figure 7: Qualitative comparison on a test image with strong scale variation. Rows: (1) CSRNet + Adaptive (baseline), (2) CSRNet + Global, (3) CSRNet + CBAM + Adaptive, (4) CSRNet + CBAM + Global.

Our result analysis supports both of our hypotheses, confirming that the proposed methods effectively address the scale variation problem in crowd estimation. The use of a global Gaussian improves supervision by assigning equal importance to all heads, allowing CSRNet to better capture both near and distant individuals without underestimating the crowd in front. Additionally, integrating CBAM enables the model to focus on important spatial and channel-wise features, helping it highlight front-facing crowds more effectively. Compared to the baseline, which often emphasizes dense, distant regions, our improved model shows better density localization and a more balanced attention across the image, confirming the effectiveness of both techniques.

6. Limitation

We were unable to find any image with ground truth head annotations that clearly distinguishes between front and back crowd regions. As a result, we generated synthetic images without ground truth density maps, which prevented us from computing objective evaluation metrics such as PSNR or SSIM. This limited our analysis to visual inspection only, making it difficult to objectively assess structural accuracy.

Additionally, our model's mean absolute error (MAE) plateaued at around 100 despite extended training. This indicates limited generalization and potential overfitting. The issue could likely be mitigated through regularization techniques such as dropout or early stopping.

7. Conclusion

We proposed improvements to CSRNet for crowd counting in scale-varying scenes by introducing a global Gaussian kernel and integrating CBAM. Our approach improves spatial focus and enables more consistent detection across varying head sizes. Experiments show better density map quality and robustness compared to the baseline. However, further studies are needed to validate these findings with more diverse datasets and precise annotations. Future work may also explore better regularization and adaptive loss strategies to enhance generalization.

References

[1] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *CVPR*, 2018.

[2] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *ECCV*, 2018.

[3] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in *CVPR*, 2016.

[4] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale Aggregation Network for Accurate and Efficient Crowd Counting," in *Lecture Notes in Computer Science*, pp. 757–773, 2018.

[5] M. A. Khan, H. Menouar, and R. Hamila, "Revisiting Crowd Counting: State-of-the-Art, Trends, and Future Perspectives," *Image and Vision Computing*, vol. 129, p. 104597, 2023.