Real-World Crowd Density Estimation and Safety Monitoring using CSRNet and MCNN

Anuj Yadav[1], Archisman[2], Kshitij Kr. Pradhan[3]
Department of Electrical Engineering
Indian Institute of Technology Bombay, India
Email: 22b3950@iitb.ac.in, 22b2405@iitb.ac.in, kshitij.kr.pradhan@iitb.ac.in

***Abstract*—Deep convolutional neural networks such as MCNN and CSRNet achieve competitive results on curated crowd counting benchmarks, but their behaviour in unconstrained real-world environments is poorly understood. This work empirically studies the generalization ability of these models when transferred from the ShanghaiTech Part A dataset to a newly collected, manually annotated real-world dataset containing low-light street markets, railway stations, and outdoor scenes. We first reproduce MCNN and CSRNet training on ShanghaiTech Part A using density map supervision. We then evaluate the same models on our real-world data without fine-tuning. While the ShanghaiTech MAE is around 415–430, the real-world MAE explodes to 1940 for MCNN and 6022 for CSRNet, clearly exposing the domain gap. Qualitative analysis using best- and worst-case examples reveals consistent failure modes related to illumination, background clutter, and perspective changes. Finally, we outline a threshold-based monitoring design that raises alerts when predicted counts exceed 100 people, and we discuss trade-offs between accuracy, speed, and robustness in safety applications.**

***Index Terms*—Crowd counting, density estimation, CSRNet, MCNN, domain shift, public safety.**

I. Introduction

Crowd density estimation is a critical component in public safety and crowd management at events such as festivals, railway stations, concerts, and political rallies. Overcrowding has been responsible for numerous tragic stampedes, and early warning systems that monitor crowd build-up can potentially save lives.

Recent work in crowd counting has focused on deep convolutional architectures operating on density maps, achieving strong results on academic datasets such as ShanghaiTech [1], [2]. However, these datasets are collected under relatively controlled imaging conditions: limited camera viewpoints, moderate illumination variability, and curated scenes. When such models are deployed in the wild, they must cope with low-light environments, motion blur, non-standard viewpoints, and strong background clutter.

This project investigates how two classic models—Multi-Column CNN (MCNN) [1] and CSRNet [2]—behave when transferred from ShanghaiTech Part A to a real-world dataset collected from CCTV-style and online video sources. Our contributions are:

- Reproduction of MCNN and CSRNet training on ShanghaiTech Part A using a simple, transparent pipeline on Kaggle.
- Construction and manual annotation of a small yet diverse real-world dataset with challenging lighting and occlusion conditions.
- Quantitative domain-transfer evaluation showing an order-of-magnitude degradation in MAE/RMSE when moving from ShanghaiTech to real-world data.
- Qualitative analysis of failure modes using best and worst examples for both MCNN and CSRNet, highlighting where and why these models break.
- A threshold-based monitoring design that flags frames as *SAFE*, *WARNING*, or *DANGER* based on predicted counts, suitable for integration into a practical safety pipeline.

II. Related Work

Early crowd counting methods relied on detection or regression on handcrafted features. MCNN [1] introduced the idea of a multi-column architecture with different receptive fields to handle scale variation, while CSRNet [2] replaced the fully connected layers of VGG with dilated convolutions to enlarge the receptive field.

Subsequent work explored lightweight models for edge devices [3], domain adaptation, and more recently diffusion and transformer-based approaches [4]. Most of these studies, however, still evaluate on benchmark datasets that share similar style and acquisition conditions. In contrast, our emphasis is on *deployment realism*: we explicitly test ShanghaiTech-trained models on scenes from Indian street markets, railway platforms, and beaches, and quantify how badly they fail.

III. Datasets

A. ShanghaiTech Part A

ShanghaiTech Part A consists of 482 congested images collected from the Internet [1]. Each person is annotated by a head location. Following common practice, we transform point annotations into continuous density maps using 2D Gaussian kernels. For simplicity and reproducibility, we use a fixed standard deviation $\sigma = 4$ pixels for all points and do not use geometry-adaptive kernels.

We adopt the official train/test split. All images are resized to 480×640 before being fed to the network. Ground truth density maps are generated in the original resolution and then bilinearly downsampled to match the network output resolution during training.

B. Real-World Dataset

To emulate realistic deployment scenarios, we curated a new dataset from publicly available video footage and CCTV-style recordings. The scenes include:

- Night-time street markets with strong point light sources, specular reflections, and significant occlusion.
- Crowded suburban railway platforms viewed from overhead bridges.
- Outdoor beaches captured from elevated viewpoints.

From these videos we extracted still frames and manually annotated head locations using a simple point-based tool. The annotated subset contains approximately 50 frames, each with between 20 and 150 people. We convert these point annotations into density maps using the same fixed-σ Gaussian procedure as for ShanghaiTech, ensuring that the integral of the density map equals the head count. This dataset intentionally violates many assumptions of ShanghaiTech: illumination, background texture, and scene composition are drastically different.

IV. Models and Training

A. MCNN

The MCNN architecture [1] processes the RGB input image using three parallel branches with convolutional kernels of size 9×9, 7×7, and 5×5, respectively. Each branch consists of several convolutional and max-pooling layers, followed by a concatenation across channels and a 1×1 convolution to produce a single-channel density map. No activation is applied to the output layer to allow arbitrary positive densities.

B. CSRNet

CSRNet [2] uses the convolutional layers of VGG16 with batch normalization as a frontend, truncated after the fourth pooling layer. The backend replaces fully connected layers with a series of dilated convolutions to enlarge the receptive field without reducing spatial resolution. A final 1×1 convolution produces a single-channel density map.

We initialise the VGG frontend with ImageNet-pretrained weights and train all layers jointly.

C. Training Setup

All experiments were implemented in PyTorch and run on Kaggle GPUs. Images are resized to 480×640 and normalised using ImageNet statistics. We use mean squared error (MSE) between the predicted and ground truth density maps as the loss function. During training, if the spatial dimensions of the prediction and ground truth do not match, the ground truth is bilinearly resized to match the prediction.

For MCNN, we use the Adam optimizer with a learning rate of 10^{-4}; for CSRNet, we use 10^{-5}. Both models are trained for 50 epochs with a batch size of 1. No sophisticated data augmentation is used; this intentionally simple setting helps isolate the effect of domain shift rather than hyperparameter tuning.

TABLE I: Performance on ShanghaiTech Part A (test set).

Model	MAE	RMSE
MCNN (ours)	414.95	540.18
CSRNet (ours)	429.40	554.58

TABLE II: Zero-shot performance on the annotated real-world dataset.

Model	MAE	RMSE
MCNN	1940.49	3486.22
CSRNet	6022.32	11548.18

V. Evaluation Protocol

We evaluate model performance using the standard metrics:

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}, \tag{2}$$

where y_i is the ground-truth count for image i (the sum of its density map), and $\hat{y}_i$ is the predicted count. We first evaluate on the ShanghaiTech Part A test set to verify that training is stable, and then perform zero-shot evaluation on the annotated real-world dataset without any fine-tuning.

VI. Quantitative Results

A. ShanghaiTech Part A

Table I summarises our reproduction results on ShanghaiTech Part A. They are significantly weaker than the original reported numbers (approximately 68 MAE), which we attribute to our simplified training configuration and lack of geometry-adaptive kernels. However, both models converge to stable behaviour and are therefore suitable as baselines for domain-transfer analysis.

B. Real-World Dataset

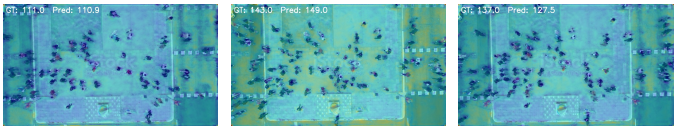
Table II shows the zero-shot performance of the ShanghaiTech-trained models on our real-world dataset. Both models exhibit a dramatic increase in error, with CSRNet degrading more severely than MCNN.

Relative to ShanghaiTech Part A, MCNN's MAE increases by approximately $4.7\times$, while CSRNet's MAE increases by roughly $14\times$. This strongly confirms that both models are highly sensitive to domain shift and that CSRNet, despite being stronger on the benchmark in prior work, is more brittle in our challenging deployment setting.

VII. Qualitative Analysis

A. MCNN Best-Case Examples

Figure 1 shows three successful predictions from MCNN on our real-world dataset. These correspond to aerial views of a pedestrian crossing with relatively uniform illumination and



GT: 111, Pred: 110.9 GT: 143, Pred: 149.0 GT: 137, Pred: 127.5

Fig. 1: MCNN best-case predictions on real-world data. The model performs well in top-view scenes with moderate density, clear separation between individuals, and limited clutter.

GT: 114, Pred: 12290 GT: 89, Pred: 12196.5 GT: 61, Pred: 11933.0

Fig. 2: MCNN worst-case predictions in night-time market scenes. Strong point lights, reflections, and textured stalls are misinterpreted as dense crowds.

limited background clutter. The ground-truth (GT) counts and predicted (Pred) counts are overlaid on the images.

In these cases, the MCNN density map aligns reasonably well with the pedestrian regions and ignores the pavement and background. This suggests that the model can generalise when the visual statistics are not too different from ShanghaiTech and when objects are reasonably separated.

B. MCNN Worst-Case Examples

Figure 2 illustrates three of the worst MCNN predictions, captured in night-time market scenes. Here, the model catastrophically overestimates the crowd count by two orders of magnitude.

The failure modes are clear: bright light sources, specular highlights on metal surfaces, and repetitive textures on stalls and banners are all incorrectly interpreted as "head blobs". The network has no understanding of physical scene structure and simply responds to local texture statistics.

C. CSRNet Best-Case Examples

Although CSRNet is more unstable on our dataset, it occasionally produces tolerable predictions, especially in bright outdoor scenarios with relatively low density (beach scenes). Examples are shown in Fig. 3.

Even in these best-case examples, CSRNet tends to underestimate or overestimate the count, and some predictions are even negative, reflecting instability in the learned mapping when applied to out-of-distribution images.

GT: 61, Pred: 8.1 GT: 21, Pred: -91.1 GT: 45, Pred: 91.4

Fig. 3: CSRNet examples on beach scenes. Even in these comparatively benign settings, CSRNet under- or overestimates the count and sometimes produces negative predictions.

GT: 146, Pred: 37675.3 GT: 155, Pred: 33142.7 GT: 127, Pred: 31273.1

Fig. 4: CSRNet worst-case predictions at railway stations. Sky gradients and train roofs dominate the density map, leading to extreme overestimation of crowd size.

D. CSRNet Worst-Case Examples

CSRNet's most severe failures occur in crowded railway station scenes, shown in Fig. 4. Despite the GT counts being around 130–160, the model predicts more than 30,000 people in some frames.

The dilated backend aggregates information over large receptive fields. In these railway scenes, smooth colour gradients in the sky and large textured regions such as train roofs are erroneously interpreted as extremely dense crowds, while actual heads contribute comparatively little to the response.

VIII. Monitoring and Alert Logic

Although a full real-time CCTV pipeline is beyond the scope of this project, we design a simple threshold-based monitoring logic that can be integrated into a video processing system. For a given frame, let $\hat{y}$ be the predicted count (e.g., from MCNN). We define three safety levels:

- **SAFE**: $\hat{y} < 100$,
- **WARNING**: $100 \leq \hat{y} < 150$,
- **DANGER**: $\hat{y} \geq 150$.

A global threshold of 100 people is chosen to reflect typical crowding limits for small public spaces and to clearly separate low-density from high-density scenarios. In a deployment setting, the monitoring module would process each frame of a prerecorded or live video, overlay the predicted count and

safety label on the frame, and trigger alarms when the label changes to *DANGER*.

IX. Discussion

Our experiments reveal several important observations:

- **Severe Domain Shift:** Both MCNN and CSRNet generalise poorly from ShanghaiTech to our real-world dataset. CSRNet, despite its stronger benchmark performance, is more brittle than MCNN in the presence of illumination and background changes.
- **Sensitivity to Illumination and Clutter:** Night markets with strong lights and reflections induce extreme false positives, particularly for CSRNet's dilated backend.
- **Lack of Negative Constraints:** Neither model explicitly enforces non-negativity or upper bounds in density, leading to negative predictions and extremely large positive values when the input distribution is far from training data.
- **Need for Domain Adaptation:** These results underscore the importance of domain-adaptive training, style-transfer augmentation, or self-supervised pretraining on unlabelled real-world footage before deploying models in safety-critical applications.

X. Conclusion

We presented an empirical study of two popular crowd counting models, MCNN and CSRNet, focusing on their behaviour under strong domain shift from ShanghaiTech Part A to a challenging real-world dataset. Our findings show that naive deployment of benchmark-trained models can lead to severe misestimation of crowd density, especially in low-light and highly cluttered scenes. MCNN exhibits somewhat lower error than CSRNet, but both are far from reliable in safety-critical scenarios.

Future work includes collecting a larger, more diverse annotated dataset, experimenting with domain adaptation techniques, enforcing physically motivated constraints on density maps, and implementing a fully optimised video-based monitoring system on resource-constrained hardware.

Acknowledgements

We thank the EE782 course instructors and TAs at IIT Bombay for guidance and feedback during the project.

References

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] X. Liu *et al.*, "LCDNet: Lightweight crowd density estimation," *arXiv preprint arXiv:2302.05374*, 2023.

[4] T. Zhou *et al.*, "CrowdDiff: Diffusion models for crowd counting," *arXiv preprint arXiv:2303.12790*, 2023.