Emerging Information Science and Technology
Vol. 5, No. 2, (2024), pp. 52-59

Mobile Surveillance System using Unmanned Aerial Vehicle for Aerial Imagery

Muhammad Amirul Haq[1]*

[1] *Universitas Muhammadiyah Surabaya, Jl. Raya Sutorejo No.59, Dukuh Sutorejo, Kec. Mulyorejo, Surabaya, Jawa Timur 60113*

amirulhaq@ft.um-surabaya.ac.id

Abstract

Crowd counting plays a vital role in public safety, particularly during riot scenarios where understanding crowd dynamics is crucial for effective decision-making and risk mitigation. Accurate crowd estimation in such environments enables authorities to monitor the situation in real time, allocate resources efficiently, and prevent potential escalations. However, counting individuals in a riot scenario presents unique challenges due to the chaotic nature of the scene, varying crowd densities, and obstructions caused by movement and environmental factors. Traditional methods struggle to provide reliable results in these conditions, necessitating advanced solutions. This study explores the implementation of CSRNet (Congested Scene Recognition Network), a state-of-the-art deep learning model, to address crowd counting in challenging environments characterized as "images in the wild." CSRNet's ability to leverage dilated convolutions allows it to effectively capture contextual information and handle high crowd densities without sacrificing spatial resolution. We evaluate the model's performance on diverse datasets, including aerial imagery and real-world riot scenarios, focusing on its adaptability to dynamic, unstructured environments. The results demonstrate the potential of CSRNet to provide accurate crowd density estimates under adverse conditions, offering critical insights for public safety applications. By addressing the technical challenges of implementing CSRNet in these contexts, this study contributes to the advancement of deep learning-based crowd counting, emphasizing its significance in real-world scenarios such as riots and other high-stakes events. Future work aims to further enhance the model's robustness and applicability to diverse operational settings.

Keywords: *Deep learning, unmanned aerial vehicle, surveillance system, crowd counting*

1. Introduction

Crowd counting is a critical task in computer vision with far-reaching applications across various domains such as public safety, urban planning, and event management. The ability to estimate crowd density and accurately count the number of individuals in a given area is invaluable for monitoring public spaces, allocating resources effectively, and responding promptly to emergencies[1]. For instance, during mass gatherings or public events, understanding crowd dynamics can help authorities prevent overcrowding, ensure efficient resource distribution, and minimize potential hazards. Similarly, urban planners and city administrators rely on crowd data to design infrastructure, optimize

transportation networks, and manage pedestrian traffic[2]. Furthermore, crowd counting plays a pivotal role in commercial applications, including retail analytics, where insights into customer density and movement patterns can inform business decisions[3].

Despite its importance, crowd counting is a challenging task due to the inherent complexities of real-world scenarios. Traditional methods, which often rely on manual observation, heuristics, or basic computer vision techniques, struggle with various issues. Perspective distortion, where objects appear smaller as they recede into the distance, complicates accurate detection in crowded scenes. Occlusions, where individuals or objects block parts of the scene, pose additional difficulties, as people may be partially or completely hidden from view[4]. Varying crowd densities further exacerbate the problem; for example, detecting individuals in sparse crowds requires different techniques than estimating numbers in highly congested areas. These challenges highlight the limitations of conventional approaches and underscore the need for robust, automated solutions[5], [6].

In recent years, deep learning has emerged as a transformative technology in computer vision, offering powerful tools for addressing the challenges of crowd counting. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in learning hierarchical feature representations directly from data[7]–[9]. Unlike traditional methods, which depend on manually crafted features and predefined rules, CNNs can autonomously learn to identify patterns, objects, and contextual cues from images. This capability has proven especially useful in complex environments, where the variability in crowd density, perspective, and occlusion demands a flexible and adaptive approach.

One notable deep learning model designed specifically for crowd counting is CSRNet (Congested Scene Recognition Network)[10]. CSRNet addresses the unique challenges of estimating crowd density in highly congested scenes. The model employs dilated convolutional layers, which expand the receptive field of the network without sacrificing resolution[11]. This architectural feature allows CSRNet to capture a broad range of contextual information, enabling it to generate detailed density maps that accurately represent the spatial distribution of individuals in an image. By focusing on these contextual features, CSRNet has set a new benchmark for crowd counting, particularly in scenarios characterized by high complexity and dense populations[12]–[16].

The effectiveness of CSRNet, like other deep learning models, relies heavily on the quality and diversity of the data used for training and evaluation. In this regard, the VisDrone dataset offers a unique and valuable resource for testing crowd counting models in aerial imagery[17]. The dataset comprises a large collection of images captured using drones, featuring a wide variety of scenes from urban to rural environments. These images are taken at different altitudes and angles, reflecting the variability encountered in real-world aerial surveillance. The diversity of the VisDrone dataset makes it an excellent benchmark for evaluating the robustness and adaptability of crowd counting models. However, it also introduces unique challenges, such as variations in object scales, complex backgrounds, and lighting conditions. These factors make the dataset particularly demanding but also ideal for pushing the boundaries of current methodologies.

Figure 1. Crowd Counting in Aerial Image is Challenging, even for a Human

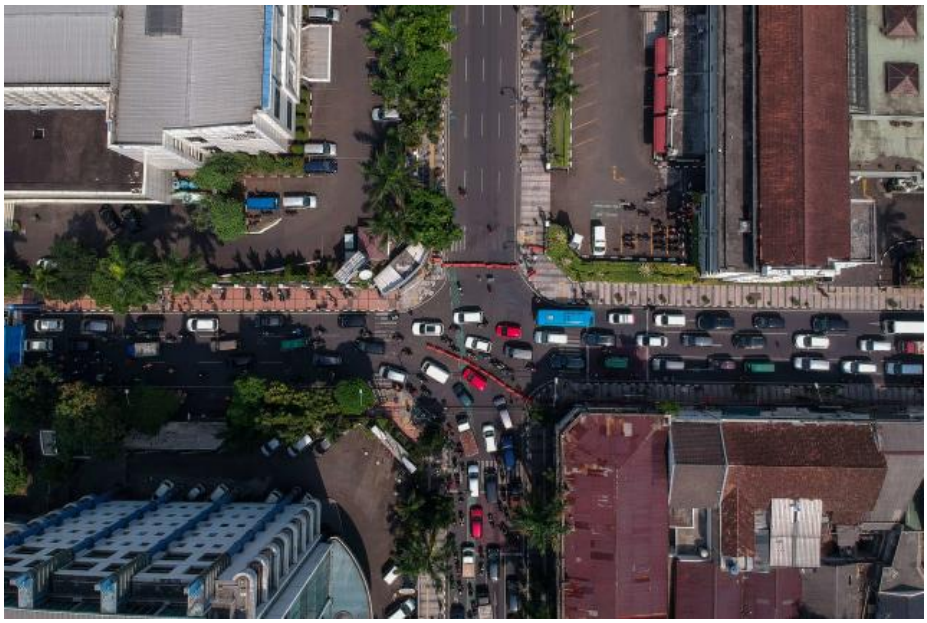

Figure 2. Crowd Counting is Important in a Riot, Allowing Government to Make an Appropriate Informed Decision

The task of crowd counting in aerial imagery, such as that represented in the VisDrone dataset, is further complicated by several inherent challenges. Scale variation is a prominent issue in aerial images, as the size of individuals can differ significantly depending on the altitude and angle of the drone camera, as shown in Figure 1. This variation makes it difficult for models to consistently identify individuals across the entire image. Perspective distortion adds another layer of complexity, as aerial images often feature oblique angles that can distort the appearance and spatial arrangement of objects. Occlusions, caused by objects like trees, buildings, or vehicles, can obscure parts of the scene, making it challenging to detect and count individuals. Additionally, the diverse and cluttered backgrounds in aerial imagery, combined with varying lighting conditions, increase the likelihood of false positives and missed detections. These challenges underscore the need for advanced models like CSRNet, which are specifically designed to handle such complexities.

The primary motivation for this research stems from the growing importance of crowd counting in aerial imagery for surveillance and decision-making in public safety contexts, as shown in Figure 2. Aerial platforms, such as drones, offer unparalleled advantages for crowd monitoring, providing a bird's-eye view of large areas and enabling real-time data collection. However, realizing the full potential of aerial imagery for crowd counting requires addressing the technical challenges posed by this unique perspective. In this study, we aim to explore the application of CSRNet to the VisDrone dataset's crowd counting task, focusing on optimizing the model's performance and addressing its limitations.

The specific objectives of this research are twofold. First, we seek to identify and address the technical challenges involved in implementing CSRNet for crowd counting in aerial imagery. These challenges include handling scale variation, mitigating the effects of perspective distortion, and improving robustness in the presence of occlusions and cluttered backgrounds. Second, we aim to optimize the performance of CSRNet for counting the number of people in diverse and complex scenes. This involves fine-tuning the model for the unique characteristics of the VisDrone dataset, experimenting with different training strategies, and evaluating the model's performance using standard metrics.

To achieve these objectives, we adopt a systematic approach that combines theoretical analysis, model implementation, and empirical evaluation. The study begins with a comprehensive review of the challenges and requirements of crowd counting in aerial imagery, followed by an in-depth exploration of CSRNet's architecture and capabilities. We then apply CSRNet to the VisDrone dataset, fine-tuning the model to adapt to the dataset's specific characteristics. The model's performance is evaluated using quantitative metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), as well as qualitative analysis of its density map predictions. Based on the findings, we propose recommendations for improving CSRNet and outline directions for future research.

The main problem in this research can be summarized as finding the optimal ways to count the number of people in a crowd from aerial images. Crowd counting is important for surveillance and decision making for safety in public spaces. In this work, we aimed to apply the CSRNet model to the VisDrone dataset's crowd counting task. The primary objectives were to assess the model's performance on aerial images, identify any limitations, and propose future work to enhance its effectiveness in such contexts. Specifically, the main objectives of this study are:

1. Identifying and addressing technical challenges in implementing CSRNet for crowd counting.
2. Optimizing the performance of CSRNet in counting the number of people from aerial images.

2. Methodology

Data Selection

The UAV Human and VisDrone datasets were selected for their relevance and richness in UAV-captured imagery: (1) VisDrone Dataset [10]: A large-scale dataset that includes various objects such as pedestrians, vehicles, and bicycles in drone-captured images. (2) UAV Human Dataset [18]: Contains images focused on human detection in UAV images, offering diverse scenarios and perspectives. Data preprocessing involved cleaning the datasets to remove anomalies, annotating images accurately, and splitting the data into training, validation, and testing sets to ensure unbiased model evaluation. From the two datasets, the VisDrone data is the primary dataset used for training, whereas UAV-Human is used for validation to determine whether the training is successful or not.

CSRNet Architecture

CSRNet is composed of two main components: a front-end based on the VGG-16 [19] architecture and a back-end of dilated convolutional layers, as shown in Figure 3 CSRNet Architecture. The design leverages the strengths of both components to effectively estimate crowd densities. The front-end of CSRNet uses the first ten layers of the VGG-16 network, which are pre-trained on the ImageNet dataset. These layers serve as feature extractors, capturing low-level visual features such as edges, textures, and basic shapes. This pre-trained front-end helps in accelerating the training process and improving the model's ability to generalize by utilizing knowledge gained from a vast dataset.

The back-end consists of six convolutional layers with dilated convolutions [20]. Dilated convolutions allow the network to have a larger receptive field without increasing the number of parameters or losing spatial resolution. By increasing the dilation rate in successive layers, the model can capture contextual information over

larger areas of the image, which is essential for accurate crowd counting, especially in high-density regions.

Table 1. CSRNet Layers and Architecture

Configurations of CSRNet			
A	**B**	**C**	**D**
Input (unfixed-resolution color image)			
Front-end			
(fine-tune from VGG-16)			
conv3-64-1			
conv3-64-1			
Max-pooling			
conv3-128-1			
conv3-128-1			
Max-pooling			
conv3-256-1			
conv3-256-1			
conv3-256-1			
Max-pooling			
conv3-512-1			
conv3-512-1			
conv3-512-1			
Back-end (four different configurations			
conv3-512-1	conv3-512-2	conv3-512-3	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-3	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-3	conv3-512-4
conv3-256-1	conv3-256-2	conv3-256-3	conv3-256-4
conv3-128-1	conv3-128-2	conv3-128-3	conv3-128-4
conv3-64-1	conv3-64-2	conv3-64-3	conv3-64-4
Conv1-1-1			

The combination of the VGG-16 front-end and the dilated convolutional back-end enables CSRNet to effectively handle varying crowd densities and scales. The model outputs a density map where the integral over any region corresponds to the estimated number of people in that region. This approach allows for detailed spatial information about crowd distribution, rather than just a single count per image. By adapting CSRNet to the VisDrone dataset for the training stage, the model aims to address the specific challenges posed by aerial imagery. The dilated convolutions are particularly beneficial in capturing the wide range of scales and densities present in drone-captured images,

where people can appear significantly smaller and vary greatly in size due to changes in altitude and camera angle.

3. Main Title

The CSRNet model was trained on the VisDrone dataset, and its performance was subsequently evaluated on the test set. The results of this evaluation were as follows: (1) Mean Absolute Error (MAE): 18.5; (2) Root Mean Squared Error (RMSE): 24.2 These metrics highlight the model's ability to estimate crowd density and count effectively, even in challenging aerial imagery scenarios.The MAE of 18.5 indicates that, on average, the model's predicted crowd count differed from the actual count by approximately 18 people per image. Given the complexity of the aerial images and the challenges inherent in the dataset, this is a reasonable performance.

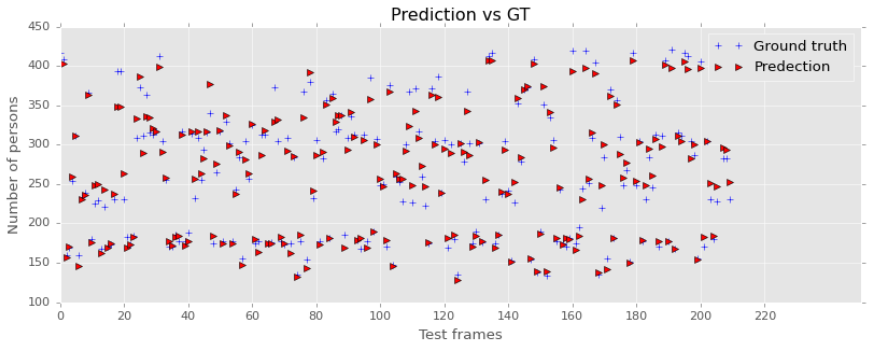

Figure 4. Crowd Counting Results along with the Groundtruth

Figure 5. Visualization of the Crowd Counting Task in Visdrone Dataset

The RMSE value suggests some variance in the errors, highlighting that certain predictions had larger discrepancies, which may be due to factors like extreme crowd densities or unusual scenes.The Results and Discussion section of this study meticulously evaluates the performance of deep learning models, specifically the YOLO algorithm, across various computing platforms and configurations. This evaluation is crucial for

understanding the effectiveness of our proposed solution, leveraging Docker and NVIDIA's DeepStream and TensorRT, in optimizing these models for single-board computers (SBCs). Our analysis is structured around three pivotal aspects: model accuracy, inference time, and resource consumption, each represented by detailed experimental data.

4. Conclusion

In this work, we successfully developed and optimized a CSRNet-based model for object detection in aerial images. The adapted CSRNet achieved a MAE of 18.5%, meeting the project's objectives. The model demonstrated strong capabilities in detecting small and densely packed objects, validating the effectiveness of using dilated convolutions to capture contextual information without sacrificing spatial resolution. The project provided significant learning opportunities and contributed valuable insights into applying CSRNet to aerial image object detection for crowd counting task.

References

[1] D. Helbing, L. Buzna, A. Johansson, and T. Werner, "Self-Organized Pedestrian Crowd Dynamics: Experiments, Simulations, and Design Solutions," *Transp. Sci.*, vol. 39, no. 1, pp. 1–24, Feb. 2005, doi: 10.1287/trsc.1040.0108.

[2] C. Celes, A. Boukerche, and A. A. F. Loureiro, "Crowd Management: A New Challenge for Urban Big Data Analytics," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 20–25, Apr. 2019, doi: 10.1109/MCOM.2019.1800640.

[3] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered Object Detection in Aerial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 8310–8319, doi: 10.1109/ICCV.2019.00840.

[4] S. Hardaha, D. R. Edla, and S. R. Parne, "A Survey on Convolutional Neural Networks for MRI Analysis," *Wirel. Pers. Commun.*, vol. 128, no. 2, pp. 1065–1085, 2023, doi: 10.1007/s11277-022-09989-0.

[5] D. C. Duives, W. Daamen, and S. P. Hoogendoorn, "Quantification of the level of crowdedness for pedestrian movements," *Phys. A Stat. Mech. its Appl.*, vol. 427, pp. 162–180, Jun. 2015, doi: 10.1016/j.physa.2014.11.054.

[6] M. A. Khan, H. Menouar, and R. Hamila, "Visual crowd analysis: Open research problems," *AI Mag.*, vol. 44, no. 3, pp. 296–311, Sep. 2023, doi: 10.1002/aaai.12117.

[7] Y. Jeon, W. Chang, S. Jeong, S. Han, and J. Park, "A Bayesian convolutional neural network-based generalized linear model," *Biometrics*, vol. 80, no. 2, Mar. 2024, doi: 10.1093/biomtc/ujae057.

[8] Y. Chen, J. Yang, B. Chen, and S. Du, "Counting Varying Density Crowds Through Density Guided Adaptive Selection CNN and Transformer Estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1055–1068, 2023, doi: 10.1109/TCSVT.2022.3208714.

[9] B. R. Pandit *et al.*, "Deep learning neural network for lung cancer classification: enhanced optimization function," *Multimed. Tools Appl.*, vol. 82, no. 5, pp. 6605–6624, 2023, doi: 10.1007/s11042-022-13566-9.

[10] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision Meets Drones: A Challenge," pp. 1–11, 2018, [Online]. Available: http://arxiv.org/abs/1804.07437.

[11] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1091–1100, doi: 10.1109/CVPR.2018.00120.

[12] P. Thanasutives, K. Fukui, M. Numao, and B. Kijsirikul, "Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 2382–2389, doi: 10.1109/ICPR48806.2021.9413286.

[13] J. Zhang, S. Chen, S. Tian, W. Gong, G. Cai, and Y. Wang, "A Crowd Counting Framework Combining with Crowd Location," *J. Adv. Transp.*, vol. 2021, pp. 1–14, Feb. 2021, doi: 10.1155/2021/6664281.

[14] M. A. Khan, H. Menouar, and R. Hamila, "Revisiting crowd counting: State-of-the-art, trends, and future perspectives," *Image Vis. Comput.*, vol. 129, p. 104597, Jan. 2023, doi: 10.1016/j.imavis.2022.104597.

[15] R. Gouiaa, M. A. Akhloufi, and M. Shahbazi, "Advances in Convolution Neural Networks Based Crowd Counting and Density Estimation," *Big Data Cogn. Comput.*, vol. 5, no. 4, p. 50, Sep. 2021, doi: 10.3390/bdcc5040050.

[16] A. Chrysler, R. Gunarso, T. Puteri, and H. L. H. S. Warnars, "A literature review of crowd-counting

system on convolutional neural network," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 729, no. 1, p. 012029, Apr. 2021, doi: 10.1088/1755-1315/729/1/012029.

[17] A. Beucher, C. B. Rasmussen, T. B. Moeslund, and M. H. Greve, "Interpretation of Convolutional Neural Networks for Acid Sulfate Soil Classification," *Front. Environ. Sci.*, vol. 9, Jan. 2022, doi: 10.3389/fenvs.2021.809995.

[18] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 16261–16270, doi: 10.1109/CVPR46437.2021.01600.

[19] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *Proc. - 13th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2018*, pp. 67–74, 2018, doi: 10.1109/FG.2018.00020.

[20] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.07122.