A data-centric approach to crowd counting: synthetic dataset creation and evaluation

Olga Cherednichenko[1,*,†], Olena Yakovleva[1,2,†] and Danylo Hrechyshkin[3,†]

[1] *Bratislava University of Economics and Management, Furdekova 16, 85104 Bratislava, Slovak Republic*
[2] *SYTOSS s.r.o, Vajnorská 10645/100, 83104 Bratislava, Slovak Republic*
[3] *Swiss Re, Mythenquai 50, 8002 Zürich, Switzerland*

Abstract

Crowd counting plays a critical role in modern security systems, public safety management, and urban analytics. However, the performance of deep learning models in this domain is largely constrained by the quality and balance of available datasets. Most existing crowd datasets suffer from significant class imbalance, limited diversity in crowd density and environmental conditions, and high annotation costs. This paper presents a data-centric approach to improving crowd counting accuracy through the creation of synthetic datasets using a custom-developed plugin for the Blender editor. The proposed tool allows automated generation of labeled crowd images with controllable parameters such as crowd density, camera angle, lighting, and scene composition. We conducted comparative experiments using the P2PNet model on both real-world and synthetic datasets. The results show that synthetic data achieves comparable accuracy for low and medium crowd densities, while high-density scenes still require further refinement. The study demonstrates that synthetic datasets can effectively complement real-world data, improving model robustness and addressing dataset imbalance. The paper also provides recommendations for optimizing synthetic data generation and outlines directions for future work in enhancing realism and evaluation metrics for synthetic datasets.

Keywords

crowd counting; synthetic dataset; data-centric AI; Blender; machine learning; computer vision; dataset balance; intelligent security systems

1. Introduction

Accurate crowd counting has become a crucial component of intelligent security and surveillance systems, particularly in the context of public safety, law enforcement, and urban management. The ability to estimate the number of people in crowded environments enables authorities to detect abnormal situations, control access during large events, and analyze crowd dynamics for preventive security measures. In recent years, deep learning–based approaches have significantly advanced the accuracy of crowd counting by employing convolutional neural networks (CNNs), density map estimation, and object detection models. However, despite algorithmic progress, the performance of these models remains highly dependent on the quality and representativeness of training datasets.

Most existing datasets, such as ShanghaiTech, UCF-QNRF, and JHU-CROWD++, exhibit severe class imbalance – dominated by low-density scenes – and lack diversity in illumination, weather, and camera perspectives. Moreover, generating labeled data for high-density crowds is labor-intensive, time-consuming, and prone to human error. These limitations reduce model generalization and hinder deployment in real-world surveillance systems, especially in complex or dynamic environments.

AISSLE-2025: The International Workshop on Applied Intelligent Security Systems in Law Enforcement, October, 30–31, 2025, Vinnytsia, Ukraine

[1,*] Corresponding author.
[†] These authors contributed equally.
olga.chrednichenko@vsemba.sk (O. Cherednichenko); olena.yakovleva@vsemba.sk (O. Yakovleva); danil.grechishkin@gmail.com (D. Hrechyshkin)
0000-0002-9391-5220 (O. Cherednichenko); 0000-0002-6129-6146 (O. Yakovleva); 0009-0007-8851-9710 (D. Hrechyshkin)

This research adopts a data-centric AI perspective, emphasizing dataset improvement as a key driver for model performance. We propose an approach based on synthetic dataset generation using the Blender 3D editor, which enables controlled creation of labeled images with varying crowd densities and environmental conditions. The developed plugin automates the process of scene generation and annotation, allowing efficient dataset expansion without manual labeling.

The goal of this study is to evaluate the effectiveness of synthetic datasets for improving the training and performance of machine learning models in crowd counting.

To achieve this goal, the following objectives were defined:

1. Analyze existing real-world crowd datasets in terms of balance, diversity, and representativeness.
2. Design and implement a Blender-based plugin for automatic generation of synthetic labeled images.
3. Conduct comparative experiments on real and synthetic datasets using state-of-the-art models (e.g., P2PNet).
4. Assess the impact of synthetic data on model accuracy and propose recommendations for future dataset design.

Based on these objectives, the study addresses the following research questions:

- *RQ1*: How does dataset imbalance affect the accuracy of crowd counting models across different crowd densities?
- *RQ2*: Can synthetic data generated in Blender provide comparable performance to real datasets for low- and medium-density scenes?
- *RQ3*: What parameters of synthetic image generation most strongly influence the model's ability to generalize?

By exploring these questions, this study contributes to the development of data-centric methodologies for AI-based security systems, offering a practical approach to dataset enhancement for robust crowd analysis.

2. The state of the art

The identification of individuals within a crowd has emerged as a pivotal application in the domain of computer vision. A significant challenge confronting researchers in this domain pertains to the heterogeneity of crowd conditions, which can be characterized as either dense, heterogeneous, or mobile. It is evident that factors such as occlusion, viewing angle, and lighting levels have a substantial impact on the accuracy of counting. It is evident that these aspects exert influence over the ultimate outcomes; however, they do not represent the sole sources of complexity. It should be noted that each distinct counting method possesses its own set of advantages and limitations.

One of the earliest documented approaches to crowd counting is Herbert Jacobs' method [1]. He also formulated a basic rule for estimating crowd density: a "sparse crowd" corresponds to approximately one person per 0.9 square meters, a "dense crowd" to one person per 0.4 square meters, and a "very dense crowd" to one person per 0.2 square meters [1, 2].

Early computational approaches relied on the detection of faces or body parts, such as heads and shoulders. However, these methods achieved limited success due to issues with occlusion and environmental variability [3, 4]. Modern approaches primarily employ deep learning, particularly Convolutional Neural Networks (CNNs) and their derivatives, such as Fully Convolutional Networks (FCNs) [5].

A concept that has gained significant traction within the research community is the generation of crowd density maps, which estimate the number of people by predicting local density levels. This approach has been demonstrated to be particularly effective in high-density crowds. Recent

methods combine classical computer vision with deep learning to process images or video in real time [6].

Several well-established object detection architectures are particularly noteworthy. Faster R-CNN [7] provides accurate localization and classification of objects, including people, by combining region proposal networks with convolutional features. YOLO (You Only Look Once) [8] achieves real-time detection with high accuracy by performing classification and localization in a single forward pass. SSD (Single Shot MultiBox Detector) [9] offers a balance between speed and precision through the use of multi-scale feature maps. The effectiveness of these models, however, largely depends on the availability of large, annotated datasets is a requirement that becomes especially critical in surveillance scenarios involving real-time video processing.

Another significant approach is Crowd Density Estimation, which does not necessitate the detection of individual people, rendering it suitable for high-density scenes where occlusion is substantial [10]. The algorithm generates a density map that represents the distribution of people at each pixel of an image or video frame. The fundamental concept is to transform the detection process into a regression task, whereby the model predicts the number of individuals in each region. CNN-based models are typically utilized for this purpose, with examples including [11]:

- MCNN is a multi-column neural network that employs multiple CNN branches that are specialized for different crowd densities.
- CSRNet is a convolutional neural network that effectively handles varying crowd densities by using dilated convolution layers to process multi-scale features.

In addition to static image-based methods, video-based approaches analyze temporal sequences to track and count individuals across frames. Optical flow techniques estimate the motion of individual pixels between successive frames in order to identify moving objects. Recurrent neural networks (RNNs), and in particular long short-term memory (LSTM) models, are capable of retaining temporal dependencies and managing complex trajectories [12]. Furthermore, hybrid methods combining counting and tracking capabilities have been developed, providing more robust performance under challenging real-world conditions [13]. For instance, some recent approaches employ a pre-classification stage that assigns an image to a certain density class and then applies the most suitable counting model [14].

The development of crowd counting methods based on CNNs is undergoing rapid evolution, driven by both competitive research and the demand for enhanced accuracy. Recent models introduced in 2024 include APGCC, PSL-Net, and FGENet, each of which proposes distinct innovations [15, 16, 17]. APGCC employs auxiliary point guidance to enhance counting precision and localization, focusing on point-based detection through auxiliary cues [15]. PSL-Net introduces the pseudo square label mechanism, representing individuals as flexible spatial regions rather than points, thereby enhancing adaptability across densities [16]. FGENet places emphasis on the extraction of fine-grained features, thus allowing high levels of accuracy to be achieved in scenarios where there is an abundance of crowd density. Furthermore, it has been demonstrated that this results in improved adaptability to images that are rich in occlusion [17].

In recent years, several novel methodologies have been introduced in the field of crowd counting, emphasizing the critical role of datasets in model accuracy and generalization. In this domain, the most frequently employed datasets encompass ShanghaiTech [18], UCF-QNRF [19], JHU-CROWD++ [20], and UCF-CC-50 [21]. Of these, the ShanghaiTech dataset has been instrumental in underpinning a substantial number of experiments and model evaluations. The current state-of-the-art method, APGCC, achieves a mean absolute error (MAE) of approximately 49 people on the ShanghaiTech Part A subset [15]. As demonstrated in 1, there has been a marked improvement in the performance of the model on the ShanghaiTech dataset from 2015 to 2024.

The majority of contemporary datasets aspire to achieve balance with regard to image resolution, scene diversity, and environmental conditions. Nevertheless, a considerable number of these datasets still demonstrate imbalanced crowd density distributions and constrained

illumination variability. Specifically, the paucity of high-density or low-light scenes constrains the ability of models to generalize to real-world surveillance environments. The quality and composition of datasets have been shown to have a direct impact on the reliability of evaluation metrics such as MAE and RMSE. It has been demonstrated that higher-density scenes tend to result in larger errors, typically due to severe occlusion and overlapping individuals [22].

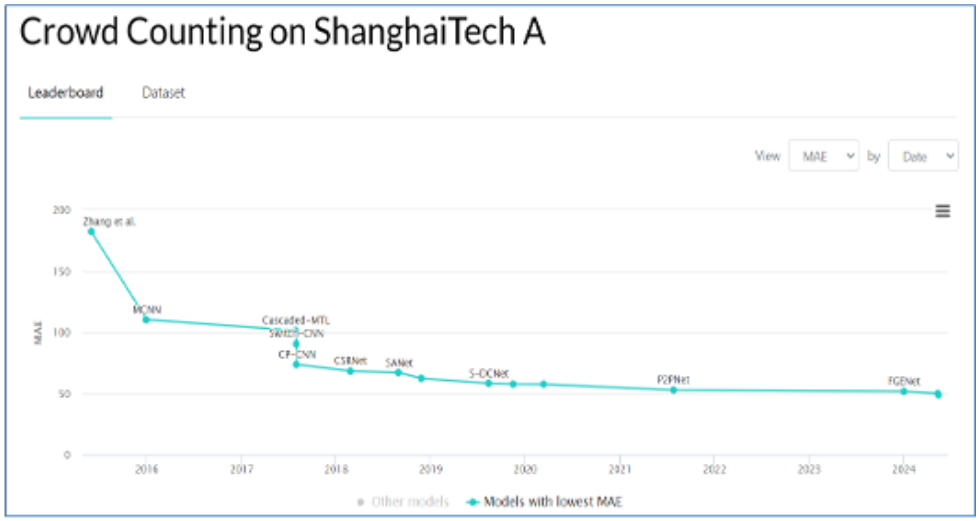

Figure 1: Counting accuracy improvement for the ShaghaiTechA dataset.

These limitations underscore a pivotal challenge: imbalanced datasets impose constraints on the capacity for enhancing model performance. In circumstances where training data is inadequate due to the presence of specific conditions, such as dense crowds, night-time lighting, or varying viewpoints, machine learning models have a tendency to overfit to dominant scenarios and subsequently underperform in underrepresented ones. The present study is predicated on the hypothesis that the balancing of datasets through synthetic image generation can enhance the performance of existing crowd counting models and provide new opportunities for improving robustness and adaptability across diverse conditions. This hypothesis forms the foundation of the proposed data-centric approach developed and evaluated in this work.

3. Methods and Materials

The present section delineates the materials, datasets, tools, and experimental methodology that were utilized in order to investigate the influence of dataset balance on the performance of crowd counting models. The proposed approach is predicated on the generation of synthetic datasets by means of a custom-developed plugin for the Blender 3D graphics editor. The plugin facilitates the automated generation of annotated crowd images, with adjustable parameters including density, lighting, camera position, and environmental context.

ShanghaiTech is a large-scale dataset for the purpose of crowd counting, containing 1,198 annotated images of crowds [18]. The dataset is divided into two parts: The part A of the collection comprises 482 images, while the second part consists of 716 images. The first part of the study is divided into two subsets: the training subset and the test subset. The training subset contains 300 images, while the test subset contains 182 images. Part B is also divided into training and test subsets, with the former comprising 400 images and the latter 316 images. Each person in the images is marked with a dot located near the centre of the head. The dataset under consideration encompasses a total of 330,165 annotated individuals. The images contained in Part A were sourced from the Internet (2), whereas those contained in Part B were collected on the bustling streets of Shanghai (3). Despite its status as a classic dataset for crowd counting network training, analysis of the images reveals that it is not balanced.

Figure 2: Random images from the ShanghaiTech Part A dataset.

Figure 3: Random images from the ShanghaiTech Part B dataset.

As demonstrated in 1, certain ranges exhibit an excess of images. For instance, the range from 50 to 99 contains 109 images, while the range from 500 to 549 contains only 3 images. The number of images containing the specified number of people was counted with quantization steps of 50 people. As demonstrated in 4, the density distribution has been observed to shift to the left.

To address dataset imbalance, this paper proposes the use of the Blender graphics editor to create artificial datasets or enrich existing real-world datasets for further model training, aiming to improve both accuracy and robustness.

Table 1
Number of images relative to specific segments

Segment	0-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499	500-549
Number of images	60	109	51	46	20	13	7	1	2	4	3

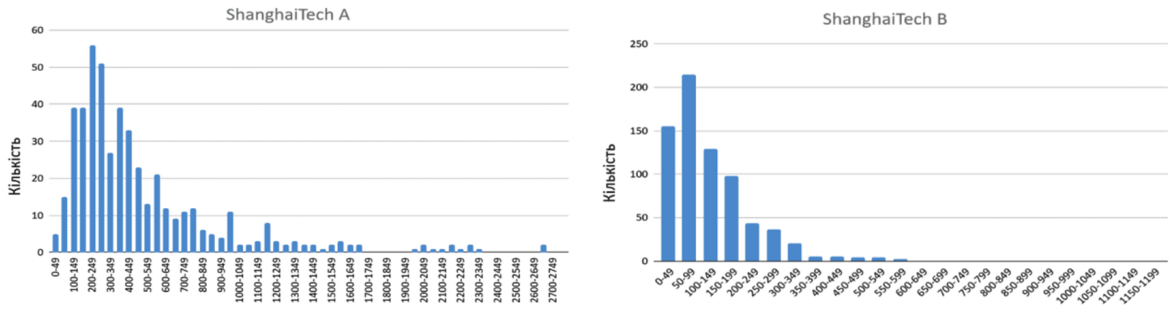

Figure 4: The ShanghaiTech dataset balance (crowd density distribution).

Blender is a powerful open-source tool for generating synthetic datasets for computer vision tasks, including crowd simulation and automatic labeling. With its ability to produce 3D scenes and its Python API for scripting, Blender enables scalable generation of crowd scenes by controlling the position, orientation, density, and other characteristics of objects. Through procedural modeling, it is possible to configure diverse lighting conditions, human poses, appearances, and dynamic motion scenarios. Blender also supports automatic annotation by generating object coordinates, bounding boxes, segmentation masks, or key points for each element in the scene – an essential feature for computer vision tasks such as detection, segmentation, and crowd counting. The use of Blender for synthetic dataset generation allows the creation of large volumes of realistic, accurately annotated data while significantly reducing the need for manual labeling.

A custom Blender plugin is developed to generate synthetic images of crowds under controlled parameters. Blender was chosen for its open-source architecture, flexibility in 3D scene creation, and integration with Python scripting for automation. The plugin allows users to specify the following parameters:

- Crowd density – the number of individuals in the scene.
- Camera angle and distance – defining the perspective and field of view.
- Lighting conditions – day, night, indoor, or weather variations.
- Environment setup – open spaces, streets, stadiums, or indoor halls.

The plugin automatically generates synthetic images and their corresponding annotation files. For each generated image, the exact 3D coordinates of individuals are projected onto a 2D plane, producing point-based labels consistent with existing real-world datasets. This ensures compatibility with standard training pipelines and enables direct comparison between real and synthetic data.

Experiments are conducted using the P2PNet architecture, a state-of-the-art regression-based model for crowd counting [23]. Training is performed on both real-world datasets and synthetic datasets generated via the Blender plugin. To assess the impact of synthetic data, two training configurations were compared:

1. Baseline – model trained on original real-world datasets only.

2. Augmented – model trained on a combination of real and synthetic data with balanced density distribution.

Model performance is evaluated using standard metrics – MAE and RMSE. For each crowd density class (low, medium, high), separate performance statistics are computed to analyze how dataset balance affects accuracy.

To validate the proposed hypothesis, results from experiments on synthetic datasets are compared with those from real-world datasets. The primary objective is to determine whether the inclusion of synthetic data improves model performance in medium- and high-density scenarios. The experiments also included visual comparisons of predicted density maps and ground truth annotations.

All experiments are performed on a workstation equipped with an NVIDIA GPU (RTX 4090, 24 GB VRAM), 128 GB RAM, and AMD Ryzen 9 processor. The software environment included Python 3.10, PyTorch 2.1, and Blender 3.6 with custom scripting modules.

The methodology combines empirical dataset analysis, procedural data generation, and comparative model evaluation. The synthetic dataset creation process provides a flexible and scalable way to produce training data that complements real datasets. This allows testing of the central hypothesis: artificially balancing the dataset improves the generalization and accuracy of existing crowd counting models, particularly under challenging conditions such as high density, occlusion, and poor lighting.

4. Experiments and Results

4.1. Experimental Design

In order to analyze how recognition will occur for the generated images, a total of more than three hundred images of different sizes, perspectives and fillings were created. In order to evaluate the results, a comparison of the metrics on the reference set (real) and the mixed set (to which artificially generated images were added) is necessary. The P2PNet model previously referenced was employed for the purpose of testing. The implementation of the local version of this network resulted in an average error of 54 people, which corresponds to [23]. An analysis of the error as a function of crowd density and the number of images in the dataset is of particular interest.

The results obtained are summarized in 2. The table illustrates that for images with low density, the error is considerably reduced, as the network demonstrates a superior ability to identify the features of individual figures. Furthermore, the number of such images is considerably higher, suggesting that the network will process images of this type more efficiently than images from other density groups. It has been demonstrated that specific groups possess an extremely limited number of instances, a factor which will have a substantial impact on the positive feature detections. It is evident that certain intervals exhibit remarkably elevated error values. For instance, the interval spanning from 1500 to 1649 demonstrates an MAE of 302.8 and an MSE of 140898, while the interval from 2100 to 2249 exhibits a similar trend. The mean absolute error (MAE) is 484.5, and the mean square error (MSE) is 270274.5. This may be attributable to an insufficient number of images for these ranges (for example, only 5 or 4 images), which hinders the training of the model. Furthermore, it is evident that certain density groups are not represented at all. Image classes such as "2400-2549" or "2700-2849" are not represented in the dataset.

The development of a Blender plugin to create a synthetic crowd dataset has the potential to enhance the efficacy of crowd counting models. This plugin has the capacity to automate the generation of scenes comprising crowds of varying density and composition, diverse lighting configurations, and temporal variations. It facilitates the acquisition of precise annotations for training models. In order to generate a substantial number of images of crowds of varying density, it is necessary for the plugin to automatically configure the scene and commence rendering.

Table 2
Accuracy by P2PNet on Shanghai A dataset

Left border	Right border	Number of images	MAE	MSE	Left border	Right border	Number of images	MAE	MSE
0	149	37	6,89	106,3	1650	1799	2	438	337005
150	299	79	16,86	585,5	1800	1949	0	-	-
300	449	63	26,17	1010,3	1950	2099	4	236,5	84664,5
450	599	34	28,44	1255,32	2100	2249	4	484,5	270274,5
600	749	26	40,07	3361,85	2250	2399	2	520,5	399442,5
750	899	17	59,11	4928,65	2400	2549	0	-	-
900	1049	12	99,92	14291,08	2550	2699	2	626	695477
1050	1199	5	71	15223	2700	2849	0	-	-
1200	1349	3	55	3597,67	2850	2999	0	-	-
1350	1499	4	128,75	26439,75	3000	3149	1	317	100489
1500	1649	5	302,8	140898					

Geometry Nodes in Blender constitutes a system that facilitates the creation and modification of the geometry of objects by means of visual programming. The configuration of geometric nodes will be stored in an object that functions as a surface of the scene on which models of people are arbitrarily placed. In the initial phase of the project, pre-created models of people were imported (see 5).

Figure 5: Models of people .

The importance of accurate camera settings in the creation of synthetic datasets cannot be overstated, given that the reliability of the scene and the quality of the resulting images are contingent on this aspect. In the context of crowd counting tasks, it is imperative to ensure that the camera is configured to reproduce a realistic perspective of the scene, particularly in cases where crowd density and lighting conditions vary. The adjustment of focal length enables the user to exercise control over the camera's angle of view. In order to simulate the human gaze, a focal length of approximately 35-50 mm is generally employed. However, in scenes encompassing substantial crowd coverage, a wider angle, such as 24 mm, may be selected. The specification of the

depth of field parameters is instrumental in the creation of a realistic scene, whereby objects in close proximity to the camera are rendered with clarity, whilst objects at greater distances may appear indistinct. This is of particular significance in the context of simulating conditions in which individuals in a crowd are distributed over a range of distances from the camera. FSpy is a free and open-source tool that facilitates the creation and adjustment of a camera in Blender according to the perspective of a photograph or image.

Head occlusion in renders is a natural phenomenon where parts of one object cover other objects in the scene, such as heads that can partially or completely cover other heads or body parts in a crowd. Following the rendering process, the coordinates of each geometric node model are translated into a camera projection. The final vector comprises the camera coordinates and the distance to the camera. Therefore, the distance value of a specific head is compared with the depth map. In the event of a discrepancy, it is determined that this head is not visible and must be filtered out.

The number of people in the frame is determined by the user, who can adjust geometric nodes to achieve this. The procedural approach to scene generation enables the user to reassemble the scene by altering specific parameters. It is evident that modifying the density parameter enables one to transition from a scene characterized by low density to one with medium density. The present study seeks to build upon the extant literature by exploring the use of crowd simulation in the context of musical performances and political demonstrations. The objective is to develop a dynamic model that can effectively capture the behaviour of large groups of people in a variety of settings. In addition to increasing the number of people on the stage plane, the user can control the distance between people, which also significantly affects the final result. Examples of such images are shown in the 6.

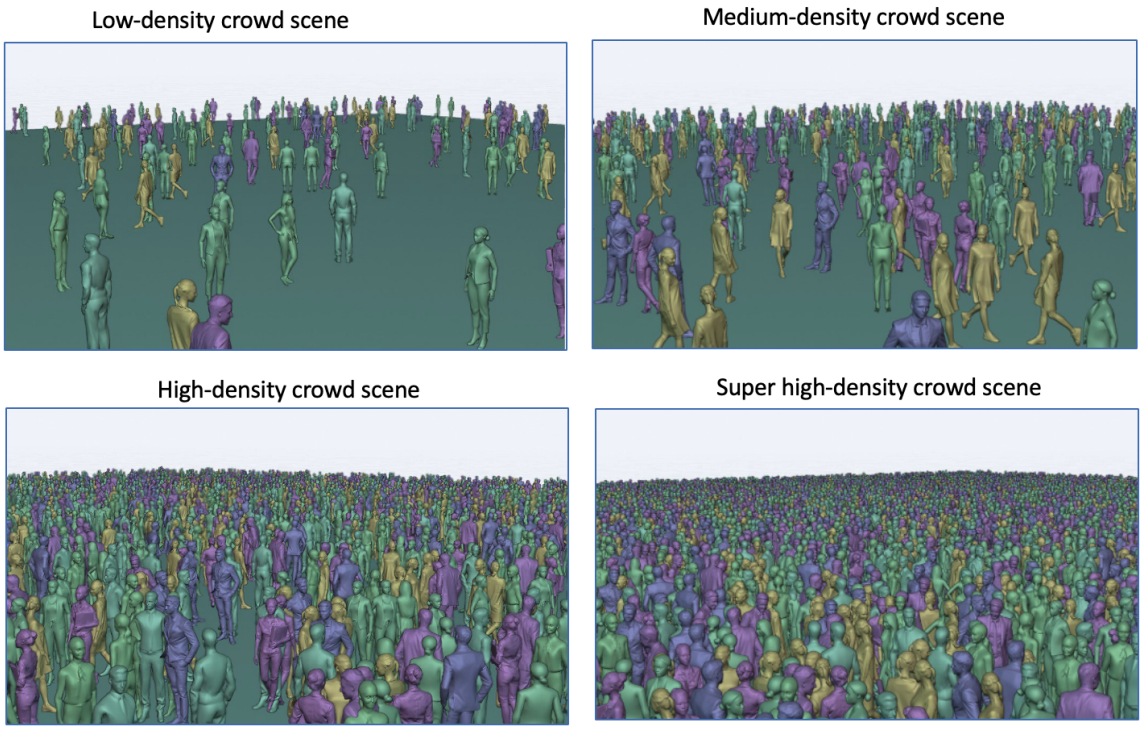

Figure 6: Creating scenes of different densities .

The creation of an artificial dataset using the developed plugin in Blender is performed by automated generation of images of a scene with a crowd and corresponding annotations for each frame. Utilizing geometric nodes and other Blender tools, a "crowd" is configured with the desired number of objects. The scene camera is set to cover the desired area with objects. Upon initiation of the rendering process, the plugin automatically saves each frame as a separate image. In addition to

rendering images, the plugin generates annotations, defined as text files containing the coordinates of objects within the camera's field of view and the number of objects present in the frame. For each segment, thirty images were generated, which should correspond to real images of this class. For the purpose of comparison, the 7 presents a series of exemplars comprising both generated and authentic images.

Figure 7: Comparison of real (left) and generated (right) images .

4.2. Experiments with Synthetic Datasets

The objective of the present series of experiments was to verify whether images produced by the proposed generation pipeline could be correctly recognized and analyzed by existing neural network models. The comparative results of the experiments are summarized in 3. In scenes characterized by low crowd density (up to approximately 450 people), the P2PNet model exhibited almost identical accuracy in both real and generated images.

Table 3
Comparison of accuracy for the ShanghaiTech A dataset and the generated dataset

Left border	Right border	MAE Shanghai A	MSE Shanghai A	MAE generated dataset	MSE generated dataset
0	149	6,89	106,3	4,73	30,13
150	299	16,86	585,5	14,2	215,47
300	449	26,17	1010,3	28,6	873,47
450	599	28,44	1255,32	45,76	2217,97
600	749	40,07	3361,85	64	9807,8

The mean absolute error (MAE) values for both datasets were comparable, thereby confirming that the synthetic data had retained sufficient realism with regard to crowd appearance, perspective, and spatial distribution. However, for high-density scenes (i.e. scenes with over 500 people), the MAE for generated images increased by approximately 30% compared to the results on ShanghaiTech Part A. This performance degradation can be attributed to severe occlusion, overlapping individuals, and imperfect representation of dense crowds in the synthetic images. These challenges suggest that errors in crowd counting models tend to increase with increasing crowd density.

These results confirm that the generated dataset can be effectively utilized for model testing and training in low- and medium-density crowd scenarios. Nevertheless, the current generation approach requires further optimization to improve realism in high-density settings. It is hypothesized that enhancing lighting simulation, introducing greater variability of human models, and refining annotation accuracy will mitigate the observed discrepancy.

The experiment demonstrates the feasibility of synthetic datasets for data augmentation and preliminary model evaluation. The findings also highlight that high-density crowd scenes remain a challenge not only for generated data but also for state-of-the-art real-world benchmarks such as ShanghaiTech Part A.

4.3. Discussion of Findings

The experimental findings confirm the central hypothesis of this research: improving dataset balance through synthetic data generation enhances model robustness and accuracy. The use of Blender for dataset generation proved effective for producing controllable, labeled images that complement real-world data. Although synthetic datasets cannot yet fully replace real data, they represent a powerful augmentation tool for addressing class imbalance and scarcity in high-density scenarios.

In summary, the experiments demonstrate that synthetic dataset generation is a viable strategy for advancing the performance of deep learning models in crowd counting. Future work should focus on improving the photorealism of generated scenes and evaluating newer architectures such as ARGCC, CLIP-EBC, or FGENet on synthetic data to explore cross-model generalization.

5. Discussion and Conclusion

This study addressed one of the central challenges in applying machine learning to intelligent security systems – the dependence of crowd counting accuracy on the quality and balance of training datasets. By adopting a data-centric AI perspective, we demonstrated that targeted improvement of datasets can significantly influence model performance without modifying model architecture.

A plugin for the Blender 3D editor was developed to generate synthetic, automatically labeled datasets for crowd counting tasks. The plugin enables control over parameters such as crowd density, camera angle, lighting, and environmental conditions, allowing researchers to simulate a wide variety of real-world scenarios. Experimental evaluation using the P2PNet model confirmed that synthetic data provides accuracy comparable to real datasets for low- and medium-density scenes, while high-density scenarios still pose challenges due to complex occlusions and limited model generalization.

The conducted research confirms that a data-centric strategy can effectively complement model-centric improvements in the field of crowd counting. While previous studies have focused primarily on developing more complex neural network architectures, this work demonstrates that the careful design and balance of datasets can yield comparable benefits. The introduction of synthetic data through Blender provides a flexible and controllable means of augmenting existing datasets, mitigating the imbalance between low- and high-density scenes. However, the experiments also revealed that the realism and diversity of the generated images remain key

factors influencing model generalization. Differences in textures, lighting realism, and human shape variation can still cause models trained on synthetic data to underperform in high-density and occluded environments.

The analysis conducted in this study provides concrete answers to the research questions formulated in the introduction. First, dataset imbalance was shown to have a significant impact on the accuracy of crowd counting models: performance drops sharply when models trained primarily on low-density scenes are applied to high-density cases. Second, the results demonstrated that synthetic data generated in Blender can achieve comparable accuracy to real-world data for low- and medium-density scenes, validating the proposed approach. Third, it was found that the most influential generation parameters include crowd density, lighting conditions, and camera perspective. Controlling these factors enables the creation of synthetic images that effectively support model generalization.

Future research will focus on improving the photorealism and variability of synthetic data. Upcoming studies will explore the integration of physics-based rendering, procedural animation of crowds, and domain adaptation techniques to reduce the synthetic-to-real gap. Another promising direction is the combination of synthetic datasets with semi-supervised or self-supervised learning frameworks, allowing models to adapt to real-world data with minimal annotation cost. Finally, the evaluation framework can be expanded beyond MAE and MSE metrics to include mAP, precision, recall, and IoU-based measures, offering a more holistic understanding of detection and localization accuracy. These efforts will contribute to building robust, scalable solutions for intelligent crowd analysis and enhance the practical deployment of AI-driven security systems in real-world environments. In summary, the proposed data-centric framework provides a practical and scalable pathway toward improving the robustness and adaptability of AI-based crowd monitoring systems — an essential step for advancing applied intelligent security technologies in real-world law enforcement and public safety contexts.

Acknowledgements

The EU NextGenerationEU partially funds the research study depicted in this paper through the Recovery and Resilience Plan for Slovakia under projects No. 09I03-03-V01-00078 and 09I03-03-V01-00115. The authors also express their gratitude to the AI R&D Department of SYTOSS s.r.o. for their valuable technical support and assistance in conducting the experiments.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI GPT-5 for text improvement (grammar and style checking). After using this tool, the authors thoroughly reviewed and edited the content and take full responsibility for the final version of the manuscript.

References

[1] A. Choi-Fitzpatrick and T. Juskauskas, “Up in the Air: Applying the Jacobs Crowd Formula to Drone Imagery,” Procedia Engineering, vol. 107, pp. 273–281, 2015, doi: 10.1016/j.proeng.2015.06.082.

[2] S. Shukla, B. Tiddeman, and H. C. Miles, “A Wide Area Multiview Static Crowd Estimation System Using UAV and 3D Training Simulator,” Remote Sensing, vol. 13, no. 14, p. 2780, Jul. 2021, doi: 10.3390/rs13142780.

[3] C. Gao, P. Li, Y. Zhang, J. Liu, and L. Wang, “People counting based on head detection combining Adaboost and CNN in crowded surveillance environment,” Neurocomputing, vol. 208, pp. 108–116, Oct. 2016, doi: 10.1016/j.neucom.2016.01.097.

[4] K. Khan, W. Albattah, R. U. Khan, A. M. Qamar, and D. Nayab, “Advances and Trends in Real Time Visual Crowd Analysis,” Sensors, vol. 20, no. 18, p. 5073, Sep. 2020, doi: 10.3390/s20185073.

[5] L. Deng, Q. Zhou, S. Wang, J. M. Górriz, and Y. Zhang, “Deep learning in crowd counting: A survey,” CAAI Transactions on Intelligence Technology, vol. 9, no. 5, pp. 1043–1077, Oct. 2024, doi: 10.1049/cit2.12241.
[6] R. Wang, Y. Hao, Y. Miao, L. Hu, and M. Chen, “RT3C: Real-Time Crowd Counting in Multi-Scene Video Streams via Cloud-Edge-Device Collaboration,” IEEE Transactions on Services Computing, vol. 17, no. 4, pp. 1739–1752, Jul. 2024, doi: 10.1109/TSC.2024.3377156.
[7] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.
[9] W. Liu et al., “SSD: Single Shot MultiBox Detector,” 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.
[10] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, “A survey of crowd counting and density estimation based on convolutional neural network,” Neurocomputing, vol. 472, pp. 224–251, Feb. 2022, doi: 10.1016/j.neucom.2021.02.103.
[11] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, “A survey of deep learning methods for density estimation and crowd counting,” Vicinagearth, vol. 2, no. 1, p. 2, Feb. 2025, doi: 10.1007/s44336-024-00011-8.
[12] M. A. Hossain, K. Cannons, D. Jang, F. Cuzzolin, and Z. Xu, “Video-Based Crowd Counting Using a Multi-scale Optical Flow Pyramid Network,” 2021, pp. 3–20. doi: 10.1007/978-3-030-69541-5_1.
[13] F. Abdullah, Y. Y. Ghadi, M. Gochoo, A. Jalal, and K. Kim, “Multi-Person Tracking and Crowd Behavior Detection via Particles Gradient Motion Descriptor and Improved Entropy Classifier,” Entropy, vol. 23, no. 5, p. 628, May 2021, doi: 10.3390/e23050628.
[14] A. N. Alhawsawi, S. D. Khan, and F. Ur Rehman, “Crowd Counting in Diverse Environments Using a Deep Routing Mechanism Informed by Crowd Density Levels,” Information, vol. 15, no. 5, p. 275, May 2024, doi: 10.3390/info15050275.
[15] I.-H. Chen, W.-T. Chen, Y.-W. Liu, M.-H. Yang, and S.-Y. Kuo, “Improving Point-based Crowd Counting and Localization Based on Auxiliary Point Guidance,” May 2024. doi: 10.48550/arXiv.2405.10589.
[16] J. Ryu and K. Song, “Crowd Counting and Individual Localization Using Pseudo Square Label,” IEEE Access, vol. 12, pp. 68160–68170, 2024, doi: 10.1109/ACCESS.2024.3400310.
[17] H.-Y. Ma, L. Zhang, and X.-Y. Wei, “FGENet: Fine-Grained Extraction Network for Congested Crowd Counting,” Jan. 2024. doi: 10.48550/arXiv.2401.01208.
[18] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.
[19] H. Idrees et al., “Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds,” 2018, pp. 544–559. doi: 10.1007/978-3-030-01216-8_33.
[20] V. A. Sindagi, R. Yasarla, and V. M. Patel, “JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method,” Nov. 2020. doi: 10.48550/arXiv.2004.03597.
[21] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source Multi-scale Counting in Extremely Dense Crowd Images,” in 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Jun. 2013, pp. 2547–2554. doi: 10.1109/CVPR.2013.329.
[22] R. Perko, M. Klopschitz, A. Almer, and P. M. Roth, “Critical Aspects of Person Counting and Density Estimation,” Journal of Imaging, vol. 7, no. 2, p. 21, Jan. 2021, doi: 10.3390/jimaging7020021.
[23] Q. Song et al., “Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework,” Aug. 2021. doi: 10.48550/arXiv.2107.12746.