Buletin Ilmiah Sarjana Teknik Elektro
Vol. 7, No. 3, September 2025, pp. 657-667
ISSN: 2685-9572, DOI: 10.12928/biste.v7i3.14391

Accurate Crowd Counting Using an Enhanced LCDANet with Multi-Scale Attention Modules

Nurmukhammed Abeuov [1], Daniyar Absatov [1], Yelnur Mutaliyev [2,4], Azamat Serek [1,3]
[1] School of Information Technologies and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan
[2] Institute of Information and Computational Technologies, Satbayev University, Almaty, Kazakhstan
[3] School of Digital Technologies, Narxoz University, Almaty, Kazakhstan
[4] Department of Computer Science, SDU University, Kaskelen, Kazakhstan

ARTICLE INFORMATION

Article History:

Received 01 August 2025
Revised 23 September 2025
Accepted 16 October 2025

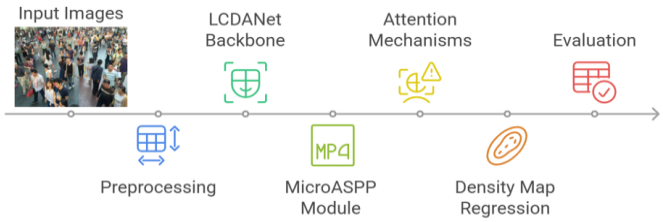
Keywords:

Crowd Counting;
Density Estimation;
MicroASPP;
Attention Mechanismss;
Inference of Crowd

Corresponding Author:

Daniyar Absatov,
School of Information Technologies and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan.
Email: da_absatov@kbtu.kz

ABSTRACT



Accurate crowd counting remains a challenging task due to occlusion, scale variation, and complex scene layouts. This study proposes ME-LCDANet, an enhanced deep learning framework built upon the LCDANet backbone, integrating multi-scale feature extraction via Micro Atrous Spatial Pyramid Pooling (MicroASPP) and attention refinement using CBAMLite modules. A preprocessing pipeline with Gaussian-based density maps, synchronized augmentations, and a dual-objective loss function combining density and count supervision supports effective training and generalization. Experimental evaluation on the ShanghaiTech Part B dataset demonstrates a Mean Absolute Error (MAE) of 11.50 (95% CI: 10.20–12.91) and a Root Mean Squared Error (RMSE) of 11.54 (95% CI: 10.26–12.99). Training dynamics indicate steadily declining loss and reduced validation MAE, while gradient norm analysis suggests reliable convergence. Comparative results show that, although CSRNet and SaNet achieve slightly lower MAE, ME-LCDANet attains a notably reduced RMSE, reflecting robustness against large prediction deviations. While the study focuses on a single benchmark dataset, the proposed architecture offers a promising approach for robust crowd counting in diverse scenarios.

Document Citation:

N. Abeuov, D. Absatov, Y. Mutaliyev, and A. Serek, "Accurate Crowd Counting Using an Enhanced LCDANet with Multi-Scale Attention Modules" *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 7, no. 3, pp. 657-667, 2025, DOI: 10.12928/biste.v7i3.14391.

1. INTRODUCTION

Accurate crowd counting and density estimation represent some of the most challenging tasks in computer vision due to the complexity of real-world scenes and the broad range of applications in which they play a critical role [1]–[3]. These tasks are vital for public safety [4]–[6], urban planning [7]–[9], intelligent transportation [10]–[12], and event management [13]–[15]. For example, reliable crowd monitoring can help authorities prevent accidents, regulate pedestrian flow, and ensure safe conditions during large gatherings [16]–[18]. Similarly, understanding patterns of human presence in public spaces assists in designing urban infrastructure and optimizing resource allocation [19]–[21]. The challenges of crowd counting stem from a variety of factors, including severe occlusion, variations in scale, perspective distortions, and cluttered backgrounds [22]–[24]. Errors in detecting small or partially visible individuals can lead to significant deviations in count accuracy, particularly in scenes with complex layouts. Traditional detection-based approaches often struggle with such conditions, missing small or occluded persons, while regression-based methods, which directly map features to global counts, frequently fail to preserve spatial precision [25]–[27]. As a result, crowd counting demands architectures that are both sensitive to fine-grained local features and capable of leveraging broader contextual information.

In recent years, deep learning has achieved remarkable success across a wide range of computer vision tasks, including object recognition, segmentation, image-to-image regression, and connections with natural language processing tasks [28]–[30]. In particular, convolutional neural networks (CNNs) have driven substantial progress in crowd counting by enabling end-to-end learning of density maps directly from images [31]–[33]. CNN-based approaches automatically extract hierarchical features that capture both local and global patterns [34]–[36], allowing them to detect subtle cues such as texture, shape, and spatial arrangement that indicate the presence of individuals, even in partially occluded or cluttered settings [37]–[39]. Modern architectures further improve performance through multi-scale feature fusion, which is essential because individuals may vary significantly in scale due to distance from the camera or perspective distortion [40]. By combining features extracted at multiple resolutions, networks can detect both small, distant individuals and larger, closer ones within the same scene [41]. In parallel, attention mechanisms, originally popularized in natural language processing [42]–[44], have been adapted to computer vision tasks to enhance feature selectivity. Attention modules help networks suppress irrelevant background while emphasizing semantically meaningful regions [45]. In the context of crowd counting, attention-guided processing allows more precise identification of individuals, improving both density map quality and total count accuracy [46].

Despite these advances, achieving high precision while maintaining computational efficiency remains a central challenge [47]–[49]. Existing models often rely on deep, complex backbones, which increase computational cost and slow inference [50]–[52], or prioritize efficiency at the expense of contextual awareness, leading to degraded performance in real-world conditions [53]. For practical deployment in real-time monitoring applications, a solution must therefore balance efficiency with strong contextual reasoning [54],[56]-[58]. In this work, we introduce MicroASPP-Enhanced LCDANet, a novel architecture designed to improve crowd counting and density estimation. LCDANet is a lightweight convolutional neural network originally proposed for crowd counting, which emphasizes multi-scale contextual aggregation with reduced computational complexity [55]. The aim of this research is to develop an efficient and accurate framework that balances contextual awareness, robustness to occlusion, and computational efficiency. To achieve this, we extend the LCDANet backbone with lightweight modules and evaluate the approach on the ShanghaiTech Part B dataset. The main contributions of this work are as follows: (i) the development of an enhanced model architecture designed for multi-scale feature learning and attention refinement, (ii) the establishment of a reliable preprocessing and training pipeline for density estimation, and (iii) a comprehensive evaluation using both quantitative and qualitative analyses to demonstrate accuracy, robustness, and interpretability.

The aim of this research is to develop an efficient and accurate deep learning framework for crowd counting and density estimation using the ShanghaiTech Part B dataset. The proposed system leverages a MicroASPP-enhanced LCDANet with CBAMLite attention modules to balance multi-scale contextual awareness, robustness to occlusion, and computational efficiency. Objectives:

1. To construct a reliable preprocessing pipeline for the ShanghaiTech Part B dataset, including Gaussian-based density map generation, resize operations that preserve person counts, and synchronized image-density augmentations.
2. To design and implement an enhanced LCDANet architecture
3. To train the model using a dual-objective loss function, combining pixel-wise density map loss with count-based supervision, optimized with Adam and cosine annealing scheduling under mixed precision training.

4. To evaluate the trained model on validation and test splits of the ShanghaiTech Part B dataset using standard metrics (MAE, RMSE), along with bootstrap confidence intervals for robust statistical assessment.
5. To analyze qualitative results by visualizing predicted density maps and comparing them with ground truth, thereby validating the interpretability and reliability of the model's predictions in sparse crowd scenarios.

The scientific novelty of this study lies in the integration of MicroASPP and CBAMLite into the LCDANet framework for crowd counting. We hypothesize that combining lightweight multi-scale contextual aggregation with compact attention refinement will improve both density map accuracy and robustness to challenging conditions such as occlusion and perspective distortion, while maintaining computational efficiency. To test this hypothesis, we evaluate the proposed MicroASPP-Enhanced LCDANet on the widely used ShanghaiTech Part B benchmark. Our results demonstrate that the model achieves state-of-the-art performance in terms of both density map quality and counting accuracy, confirming the effectiveness of the proposed architectural enhancements.

2. LITERATURE REVIEW

2.1. Traditional Approaches to Crowd Counting

Early methods for crowd counting were dominated by detection-based and regression-based approaches. Detection-based methods attempted to identify and count each individual in the scene using handcrafted features such as Haar wavelets, HOG descriptors, and edge features [58]. While effective for sparse crowds, these approaches struggled in dense or occluded environments due to overlapping pedestrians. Regression-based methods emerged as an alternative, mapping low-level image features directly to global crowd counts [59]. Although more robust against occlusion, these methods discarded spatial information, limiting their ability to generate accurate density maps.

2.2. Density Map Estimation and Deep Learning Advances

The introduction of density map estimation significantly improved performance in crowd counting tasks. This approach not only predicted the total count but also provided spatial distributions of individuals, enabling more detailed analysis. The advent of Convolutional Neural Networks (CNNs) further advanced the field, with architectures such as MCNN [60] exploiting multi-column structures to extract features at different scales. Similarly, CSRNet [61] demonstrated that dilated convolutions could capture wide contextual information without significant computational overhead. These advancements highlighted the importance of multi-scale feature extraction in addressing scale variation caused by perspective distortion.

2.3. Attention Mechanisms in Crowd Counting

Attention mechanisms have been widely adopted in computer vision, inspired by their success in Natural Language Processing (NLP). In NLP, attention enables models to focus on semantically important words in a sentence, improving tasks such as translation and sentiment analysis [62]. In crowd counting, spatial and channel attention mechanisms allow networks to emphasize informative regions while suppressing irrelevant background noise. For example, SANet [63] integrates attention modules to improve the representation of highly relevant features, thereby enhancing density map quality. Recent variants, such as CBAM and its derivatives, have shown promise in balancing accuracy with computational efficiency [64].

2.4. Towards Multi-Scale Attention Architectures

Recent studies have highlighted the importance of computationally efficient models for real-world deployment in surveillance and resource-constrained environments. Architectures combining multi-scale feature extraction and attention mechanisms have emerged as a promising direction. However, many state-of-the-art models remain computationally heavy, limiting their applicability [65]. This gap motivates the development of MicroASPP-Enhanced LCDANet, which leverages multi-scale pooling and efficient attention modules (CBAMLite) to achieve accurate crowd counting without incurring excessive computational cost.

3. METHODOLOGY

This section outlines the methodological framework adopted in developing the MicroASPP-Enhanced LCDANet architecture for crowd counting. The methodology consists of four main stages: dataset preparation, baseline model selection, architectural enhancements, and model training and evaluation. Figure 1 illustrates the proposed methodology for accurate crowd counting using the MicroASPP-Enhanced LCDANet with attention mechanisms on the ShanghaiTech Part B dataset. The workflow begins with input images that

undergo preprocessing, including resizing, normalization, and data augmentation, to ensure consistency and robustness. The backbone of the proposed system is LCDANet (Lightweight Contextual Dilated Attention Network), originally designed for crowd counting tasks [55]. LCDANet employs dilated convolutions and attention mechanisms to balance computational efficiency with contextual awareness, making it a strong foundation for lightweight crowd density estimation. We extend LCDANet by integrating the Micro Atrous Spatial Pyramid Pooling (MicroASPP) module to capture multi-scale contextual features while maintaining efficiency. This enables the model to effectively represent varying head sizes across crowd scenes. To further refine feature extraction, CBAMLite attention modules are applied, highlighting the most informative regions while suppressing irrelevant background noise. The integrated features are then passed through the density map regression head, which produces high-quality density maps corresponding to the crowd distribution. Finally, the generated density maps are aggregated and evaluated to estimate the total crowd count, with performance metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) employed to validate accuracy.

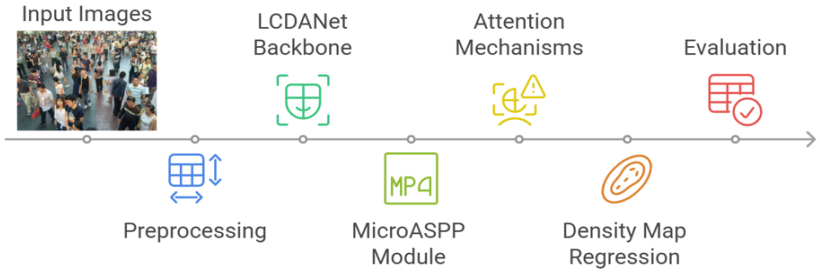


Figure 1. Proposed methodology

3.1. Dataset and Preprocessing

The ShanghaiTech Part B dataset was employed for training and evaluation. Ground-truth annotations in .mat format were converted into Gaussian density maps with $\sigma = 5.0$, ensuring that the total count of individuals is preserved. Images and density maps were resized to 384×384, and geometric transformations, including horizontal flips, were applied jointly to images and density maps to maintain alignment. Non-geometric augmentations, such as brightness and contrast adjustments and Gaussian blur, were applied to images only. Images were normalized using standard ImageNet mean and standard deviation and converted to PyTorch tensors, while density maps were converted to single-channel tensors.

3.2. Proposed Model

We introduce MicroASPP-Enhanced LCDANet (ME-LCDANet), a lightweight convolutional architecture designed for efficient and accurate crowd counting and density estimation. The backbone, LCDANet, was chosen for its ability to preserve both local detail and global context through dual-orientation feature extraction, offering competitive representational power while reducing computational cost—a key consideration for real-time applications. The network begins with a convolutional stem followed by depthwise-separable convolution blocks to extract low-level features. These features are then processed through two orientation-specific branches, which capture horizontal and vertical patterns to enhance sensitivity to scale variation and perspective distortions commonly observed in crowd scenes. Each branch incorporates a Micro Atrous Spatial Pyramid Pooling (MicroASPP) module, which aggregates multi-scale contextual information while maintaining computational efficiency. MicroASPP employs a 1×1 convolution and three 3×3 depthwise-separable convolutions with dilation rates of 1, 2, and 3, respectively, concatenates the outputs, applies a 1×1 projection, and adds a residual connection. The outputs of both branches are concatenated and passed through a fusion module comprising a 1×1 convolution, batch normalization, GELU activation, and a CBAMLite attention module. CBAMLite combines a channel-wise SE attention mechanism with spatial attention derived from both mean and max pooling, followed by a 7×7 convolution and sigmoid activation, enabling the network to suppress irrelevant background and emphasize informative regions. The final density map is produced by a convolutional decoder with a Softplus activation to ensure non-negative, smooth predictions. Optionally, the architecture can output per-pixel uncertainty through an auxiliary head.

3.3. Model Training and Evaluation

The network was trained under a dual-objective loss function that combines both pixel-wise and global supervision as illustrated in equation (1). Where D_{pred} and D_{gt} are predicted and ground-truth density maps,

and C_{pred}, C_{gt} are the corresponding crowd counts. The hyperparameter $\lambda = 0.05$ was determined empirically through preliminary experiments on the validation set. Larger values were found to overweight the count-level loss, leading to overly smoothed density maps, whereas smaller values diminished the contribution of global supervision, resulting in accurate local density patterns but less reliable total counts. The chosen value of 0.05 provided the best trade-off, ensuring spatial fidelity in the density maps while maintaining global counting accuracy.

$$L = MSE(D_{pred}, D_{gt}) + \lambda \cdot MAE(C_{pred}, C_{gt}) \quad (1)$$

Optimization was performed using the Adam optimizer with a cosine annealing learning rate scheduler. To improve training stability and efficiency, mixed precision training with gradient scaling was employed. Training and validation were conducted with batch sizes of 4 and 1, respectively, for 10 epochs. The small validation batch size was adopted to ensure accurate count evaluation per image, since averaging over larger batches can obscure sample-level errors. Additionally, hardware memory constraints during evaluation with high-resolution inputs limited the feasible validation batch size. Although the number of epochs appears relatively small, we observed rapid convergence of both training and validation losses within this range, with minimal further improvement beyond 10 epochs. Moreover, this choice balanced performance with computational resource constraints, ensuring efficient experimentation while avoiding overfitting. Extending training beyond 10 epochs produced only marginal improvements in MAE and RMSE, while substantially increasing training time. Future work may explore longer training schedules or alternative learning rate strategies to potentially further enhance performance. Performance was assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which are standard metrics in crowd counting. To provide a statistically robust assessment, bootstrap resampling (2,000 iterations) was used to compute 95% confidence intervals for both metrics on the test set.

3.4. Model Testing and Statistical Evaluation

After training, the final model parameters were stored and subsequently reloaded for evaluation on the independent test set to ensure consistency of the results. The trained model was placed in evaluation mode, thereby disabling gradient computation and ensuring deterministic inference. Predictions were generated across the entire test set, and the total crowd counts were obtained by summing the predicted density maps. These predictions were then compared against the corresponding ground truth counts to compute absolute errors and squared errors for each test sample. To provide a rigorous statistical assessment of model performance, bootstrap resampling was employed to estimate confidence intervals for both the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). This approach allows the reported metrics to be complemented with uncertainty bounds, thereby reflecting the statistical reliability of the results and mitigating the risk of overfitting to specific data samples.

Beyond numerical evaluation, several diagnostic plots were generated to provide deeper insights into the model's behavior. Scatter plots of true versus predicted counts were produced to visualize the alignment of predictions with ground truth across different crowd sizes. Histograms of prediction errors were constructed to analyze the distribution of deviations, highlighting potential bias or variance tendencies in the model. Additionally, boxplots of errors, complemented with confidence interval annotations, offered a robust visualization of prediction variability and extreme outliers. Finally, a structured summary table was compiled to present the evaluation metrics alongside their estimated confidence intervals. This tabular representation provides a concise yet comprehensive overview of the model's performance, enabling transparent comparison with alternative approaches and establishing the statistical significance of the reported results.

4. RESULTS AND DISCUSSION

Figure 2 presents the training dynamics of the proposed MicroASPP-Enhanced LCDANet model over ten epochs. The left panel illustrates the evolution of the training loss, which shows a gradual and consistent decline from approximately 0.64 at the first epoch to 0.59 at the tenth epoch. This steady reduction in loss indicates effective optimization and demonstrates that the model successfully learned to approximate the ground truth density maps without signs of divergence or overfitting within the observed training period. The right panel depicts the validation metrics, specifically Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), across the same epochs. The validation MAE exhibits a downward trajectory, decreasing from around 12.8 to 11.9 by the final epoch. This trend confirms the model's ability to generalize effectively to unseen validation data. In contrast, the validation RMSE remains relatively stable, fluctuating narrowly around 16.0–16.5 before slightly decreasing to 16.1 at the last epoch. The stability of RMSE suggests that while the model

reduced the average prediction error (as reflected in MAE), the occurrence of larger errors persisted but did not escalate, thereby maintaining robustness throughout training.

Figure 3 shows the average gradient norm values across epochs during the training process. Gradient norms provide an indication of the magnitude of updates applied to model parameters. From the plot, we observe that the gradient norms start relatively high (~166–167) and fluctuate slightly over the first few epochs. Around epoch 7–8, the gradient norms decrease sharply, reaching a minimum (~137), before increasing again towards the end of training. This trend may reflect the optimizer traversing regions of the parameter space with smaller gradients, potentially suggesting proximity to flatter regions of the loss landscape. The subsequent rise in gradient norms after epoch 8 may indicate continued parameter refinement or adjustments in the optimization trajectory. While these observations are suggestive, they are interpretative and do not constitute direct proof of flatter minima.

Table 1 showcases the quantitative evaluation results of the proposed MicroASPP-enhanced LCDANet with CBAMLite attention modules on the ShanghaiTech Part B dataset. The performance is reported using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), along with their corresponding 95% bootstrap confidence intervals. The model achieves an average MAE of 11.50 (95% CI: 10.20–12.91) and an RMSE of 11.54 (95% CI: 10.26–12.99). The confidence intervals for both metrics span approximately 2.7, which is relatively narrow compared to the mean values. This indicates that the model's predictions are consistent across different test samples, with limited variability and no extreme errors dominating the results. A lower MAE reflects that, on average, predicted counts deviate by about 11–12 individuals from the ground truth, while the low RMSE further confirms the model's stability and robustness across the dataset.

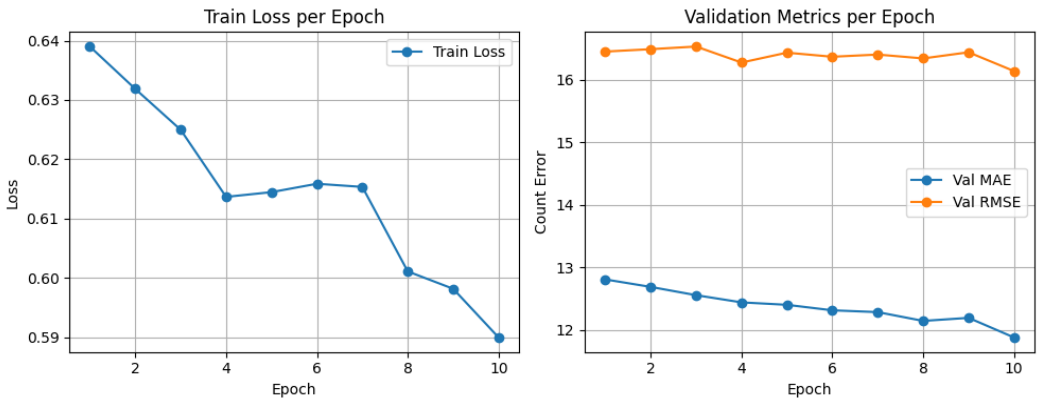


Figure 2. Train loss per epoch and validation metrics per epoch

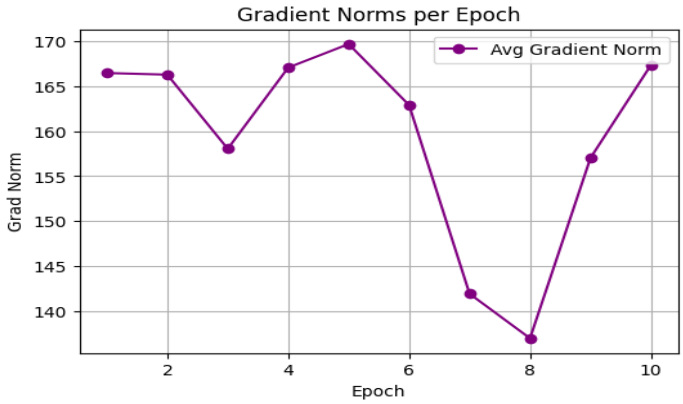


Figure 3. Average gradient norm per epochs

It is important to note that the validation RMSE trends reported in Figure 2 do not directly align with the final test RMSE reported in Table 1. This apparent discrepancy arises from two factors. First, the validation metrics were monitored throughout training using a fixed validation split, whereas the final results were obtained on the held-out test set after model selection. Second, the final model used for evaluation was selected based on the epoch yielding the best validation MAE, rather than the final epoch illustrated in Figure 2.

Consequently, the final test RMSE of 11.54 reflects the generalization performance of the best checkpoint, which is expected to outperform the intermediate validation results presented in the training dynamics. This difference is consistent with common deep learning practices, where model checkpoints selected through validation often achieve lower error on the test set than indicated by raw training curves. Table 2 showcases the relative performance of the proposed model against widely cited benchmarks. MCNN achieves the weakest performance (MAE = 26.4, RMSE = 41.3), reflecting limitations in early CNN-based architectures. CSRNet and SaNet, both advanced models, demonstrate superior accuracy with MAE/RMSE values of 10.6/16.0 and 8.4/13.6, respectively. The proposed model obtains an MAE of 11.50 and an RMSE of 11.54, which—although slightly higher in MAE than CSRNet and SaNet—exhibits a substantially lower RMSE. This indicates that the proposed architecture reduces extreme prediction errors and produces more consistent results across samples.

These findings suggest that while CSRNet and SaNet achieve lower average count errors, the proposed MicroASPP-enhanced LCDANet offers a favorable balance between accuracy and robustness. The lower RMSE highlights its ability to maintain stability across varying crowd scenarios, reducing the likelihood of extreme miscounts. This robustness is practically significant in real-world monitoring applications—such as public safety management, transportation hubs, and event crowd regulation—where occasional large prediction errors could compromise decision-making. By minimizing such deviations, the proposed model provides more reliable estimates that can be directly applied in operational settings requiring consistent crowd analysis. Nonetheless, the study is limited to the ShanghaiTech Part B dataset; future work could explore additional benchmarks, including higher-density or multi-scene datasets, as well as extending the approach to video-based temporal crowd analysis for further improving robustness and real-time applicability.

Table 1. Performance evaluation of the proposed MicroASPP-enhanced LCDANet with CBAMLite modules on the ShanghaiTech Part B dataset

Metric	Mean	95% CI Lower	95% CI Upper
MAE	11.500877	10.196053	12.906230
RMSE	11.544531	10.263619	12.989083

Table 2. Comparison of crowd counting performance on the ShanghaiTech Part B dataset across baseline models and the proposed MicroASPP-enhanced LCDANet with CBAMLite modules

Model	MAE	RMSE
MCNN [60]	26.4	41.3
CSRNet [61]	10.6	16.0
SaNet [63]	8.4	13.6
Our model	**11.500**	**11.544**

5. CONCLUSION

This study introduced a MicroASPP-enhanced LCDANet with CBAMLite attention modules for crowd counting and density estimation on the ShanghaiTech Part B dataset. The framework was supported by a robust preprocessing pipeline and a dual-objective loss function, enabling effective training and reliable prediction. Quantitative evaluation demonstrated competitive performance, with consistent and robust predictions across test samples. Training dynamics further validated effective optimization and generalization, while gradient norm analysis highlighted stable convergence behavior. In comparative evaluation, the proposed model achieved results comparable to established approaches such as CSRNet and SaNet, while exhibiting robustness against large deviations in prediction. Future work will focus on extending the framework to different crowd densities, incorporating transformer-based modules for enhanced contextual modeling, and exploring cross-dataset generalization to strengthen practical applicability in real-world scenarios. The proposed architecture contributes a reliable and adaptable approach to crowd counting that can support real-time monitoring, public safety, and urban management applications.

DECLARATION

Author Contribution

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding

This research received no external funding

Conflicts of Interest
The authors declare no conflict of interest.

REFERENCES

[1] B. Li, H. Huang, A. Zhang, P. Liu, and C. Liu, "Approaches on crowd counting and density estimation: a review," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 853–874, 2021, https://doi.org/10.1007/s10044-021-00959-z.
[2] M. Wang, X. Zhou, and Y. Chen, "A comprehensive survey of crowd density estimation and counting," *IET Image Processing*, vol. 19, no. 1, p. e13328, 2025, https://doi.org/10.1049/ipr2.13328.
[3] A. Serek, B. Amirgaliyev, R. Y. M. Li, A. Zhumadillayeva and D. Yedilkhan, "Crowd Density Estimation Using Enhanced Multi-Column Convolutional Neural Network and Adaptive Collation," in *IEEE Access*, vol. 13, pp. 146956-146972, 2025, https://doi.org/10.1109/ACCESS.2025.3597393.
[4] T. Daware and T. Dhote, "Enhancing public safety through real-time crowd density analysis and management," in *Proc. 2023 5th Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, pp. 1040–1046, 2023, https://doi.org/10.1109/ICIRCA57980.2023.10220731.
[5] M. Patidar, P. K. Bhanodia, P. K. Patidar, R. Shukla, K. Gupta, and S. Rajpoot, "Advanced crowd density estimation using hybrid CNN models for real-time public safety applications," *Library of Progress – Library Science, Information Technology & Computer*, vol. 44, no. 3, 2024, https://openurl.ebsco.com/EPDB%3Agcd%3A4%3A23637460/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A180918700&crl=c&link_origin=scholar.google.com.
[6] Y. Liu, Z. Yu, H. Cui, S. Helal, and B. Guo, "SafeCity: A heterogeneous mobile crowd sensing system for urban public safety," *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 18330–18345, 2023, https://doi.org/10.1109/JIOT.2023.3279385.
[7] R. Jiang, Z. Cai, Z. Wang, C. Yang, Z. Fan, Q. Chen, … and R. Shibasaki, "DeepCrowd: A deep model for large-scale citywide crowd density and flow prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 276–290, 2021, https://doi.org/10.1109/TKDE.2021.3077056.
[8] N. K. Saini and R. Sharma, "Deep learning approaches for crowd density estimation: A review," in *Proc. 2023 12th Int. Conf. Syst. Modeling & Advancement Res. Trends (SMART)*, pp. 83–88, 2023, https://doi.org/10.1109/SMART59791.2023.10428557.
[9] M. H. El-Didy, G. F. Hassan, S. Afifi, and A. Ismail, "Crowding between urban planning and environmental psychology: Guidelines for bridging the gap," *Open House International*, vol. 49, no. 4, pp. 670–695, 2024, https://doi.org/10.1108/OHI-06-2023-0146.
[10] D. Darsena, G. Gelli, I. Iudice, and F. Verde, "Sensing technologies for crowd management, adaptation, and information dissemination in public transportation systems: A review," *IEEE Sensors Journal*, vol. 23, no. 1, pp. 68–87, 2022, https://doi.org/10.1109/JSEN.2022.3223297.
[11] Y. Zhu, K. Ni, X. Li, A. Zaman, X. Liu, and Y. Bai, "Artificial intelligence aided crowd analytics in rail transit station," *Transp. Res. Rec.*, vol. 2678, no. 2, pp. 481–492, 2024, https://doi.org/10.1177/03611981231175156 .
[12] W. Tang, K. Liu, M. S. Shakeel, H. Wang, and W. Kang, "DDAD: detachable crowd density estimation assisted pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 1867–1878, 2022, https://doi.org/10.1109/TITS.2022.3222692.
[13] V. X. Gong, W. Daamen, A. Bozzon, and S. P. Hoogendoorn, "Counting people in the crowd using social media images for crowd management in city events," *Transportation*, vol. 48, no. 6, pp. 3085–3119, 2021, https://doi.org/10.1007/s11116-020-10159-z.
[14] O. Elharrouss, H. H. Mohammed, S. Al-Maadeed, K. Abualsaud, A. Mohamed and T. Khattab, "Crowd density estimation with a block-based density map generation," *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1-7, 2024, https://doi.org/10.1109/ISCV60512.2024.10620151.
[15] X. Zhang, Y. Sun, Q. Li, X. Li, and X. Shi, "Crowd density estimation and mapping method based on surveillance video and GIS," *ISPRS Int. J. Geo-Inf.*, vol. 12, no. 2, p. 56, 2023, https://doi.org/10.3390/ijgi12020056.
[16] M. Chengo, J. Bitok, and S. W. Maingi, "Crowd management, risk assessment strategies and sports tourism events in Nairobi County, Kenya," *J. Hospitality Tourism Manage.*, vol. 7, no. 1, pp. 46–66, 2024, https://doi.org/10.53819/81018102t4253.
[17] S. Huang, J. Ji, Y. Wang, W. Li, and Y. Zheng, "A machine vision-based method for crowd density estimation and evacuation simulation," *Safety Science*, vol. 167, p. 106285, 2023, https://doi.org/10.1016/j.ssci.2023.106285.
[18] A. M. Alasmari, N. S. Farooqi, and Y. A. Alotaibi, "Recent trends in crowd management using deep learning techniques: a systematic literature review," *J. Umm Al-Qura Univ. Eng. Archit.*, pp. 1–29, 2024, https://doi.org/10.1007/s43995-024-00071-3.
[19] T. Zhang, J. Yuan, Y. C. Chen, and W. Jia, "Self-learning soft computing algorithms for prediction machines of estimating crowd density," *Appl. Soft Comput.*, vol. 105, p. 107240, 2021, https://doi.org/10.1016/j.asoc.2021.107240.
[20] M. Fiandeiro, T. T. Nguyen, H. Wong, and E. B. Hsu, "Modernized crowd counting strategies for mass gatherings—A review," *J. Acute Med.*, vol. 13, no. 1, p. 4, 2023, https://doi.org/10.6705/j.jacme.202303_13(1).0002.
[21] S. Goel, D. Koundal, and R. Nijhawan, "Learning models in crowd analysis: A review," *Arch. Comput. Methods Eng.*, vol. 32, no. 2, pp. 943–961, 2025, https://doi.org/10.1007/s11831-024-10151-1.

[22] Y. C. Li, R. S. Jia, Y. X. Hu, and H. M. Sun, "A lightweight dense crowd density estimation network for efficient compression models," *Expert Syst. Appl.*, vol. 238, p. 122069, 2024, https://doi.org/10.1016/j.eswa.2023.122069.
[23] J. P. Singh, M. Kumar, A. Arya, and B. Badmera, "Scientific exploration for density estimation and crowd counting of crowded scene," in *J. Phys.: Conf. Ser.*, vol. 1947, no. 1, p. 012019, 2021, https://doi.org/10.1088/1742-6596/1947/1/012019.
[24] Y. Ranasinghe, N. G. Nair, W. G. C. Bandara, and V. M. Patel, "CrowdDiff: Multi-hypothesis crowd density estimation using diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12809–12819, 204, https://doi.org/10.1109/CVPR52733.2024.01217.
[25] A. A. Assefa, W. Tian, N. W. Hundera, and M. U. Aftab, "Crowd density estimation in spatial and temporal distortion environment using parallel multi-size receptive fields and stack ensemble meta-learning," *Symmetry*, vol. 14, no. 10, p. 2159, 2022, https://doi.org/10.3390/sym14102159.
[26] Y. C. Li, R. S. Jia, Y. X. Hu, D. N. Han, and H. M. Sun, "Crowd density estimation based on multi scale features fusion network with reverse attention mechanism," *Appl. Intell.*, vol. 52, no. 11, pp. 13097–13113, 2022, https://doi.org/10.1007/s10489-022-03187-y.
[27] M. Wang, X. Zhou, and Y. Chen, "A comprehensive survey of crowd density estimation and counting," *IET Image Process.*, vol. 19, no. 1, p. e13328, 2025, https://doi.org/10.1049/ipr2.13328.
[28] X. Zhang, Y. Sun, Q. Li, X. Li, and X. Shi, "Crowd density estimation and mapping method based on surveillance video and GIS," *ISPRS Int. J. Geo-Inf.*, vol. 12, no. 2, p. 56, 2023, https://doi.org/10.3390/ijgi12020056.
[29] 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, 2021, https://doi.org/10.3390/bdcc5040050.
[30] G. Yang and D. Zhu, "Survey on algorithms of people counting in dense crowd and crowd density estimation," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13637–13648, 2023, https://doi.org/10.1007/s11042-022-13957-y.
[31] L. Zholshiyeva, T. Zhukabayeba, A. Serek, R. Duisenbek, M. Berdieva, and N. Shapay, "Deep learning-based continuous sign language recognition," *Journal of Robotics and Control (JRC)*, vol. 6, no. 3, pp. 1106–1118, 2025, https://doi.org/10.18196/jrc.v6i1.23879.
[32] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024, https://doi.org/10.1007/s10462-024-10721-6.
[33] L. Zholshiyeva, T. Zhukabayeva, D. Baumuratova, and A. Serek, "Design of QazSL sign language recognition system for physically impaired individuals," *Journal of Robotics and Control (JRC)*, vol. 6, no. 1, pp. 191–201, 2025, https://doi.org/10.18196/jrc.v6i1.23879.
[34] 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, https://doi.org/10.1016/j.imavis.2022.104597.
[35] L. Dong, H. Zhang, K. Yang, D. Zhou, J. Shi, and J. Ma, "Crowd counting by using top-k relations: A mixed ground-truth CNN framework," *IEEE Transactions on Consumer Electronics*, vol. 68, no. 3, pp. 307–316, 2022, https://doi.org/10.1109/TCE.2022.3190384.
[36] S. Zhang, W. Wang, W. Zhao, L. Wang, and Q. Li, "A cross-modal crowd counting method combining CNN and cross-modal transformer," *Image and Vision Computing*, vol. 129, p. 104592, 2023, https://doi.org/10.1016/j.imavis.2022.104592.
[37] P. Purwono, A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. ul Haq, "Understanding of convolutional neural network (CNN): A review," *International Journal of Robotics and Control Systems*, vol. 2, no. 4, pp. 739–748, 2022, https://doi.org/10.31763/ijrcs.v2i4.888.
[38] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, *et al.*, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, 2021, https://doi.org/10.3390/electronics10202470.
[39] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, *et al.*, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, 2021, https://doi.org/10.1186/s40537-021-00444-8.
[40] J. Gupta, S. Pathak, and G. Kumar, "Deep learning (CNN) and transfer learning: A review," in *Journal of Physics: Conference Series*, vol. 2273, no. 1, p. 012029, 2022, https://doi.org/10.1088/1742-6596/2273/1/012029.
[41] J. Lu, L. Tan, and H. Jiang, "Review on convolutional neural network (CNN) applied to plant leaf disease classification," *Agriculture*, vol. 11, no. 8, p. 707, 2021, https://doi.org/10.3390/agriculture11080707.
[42] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, *et al.*, "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023, https://doi.org/10.3390/su15075930.
[43] 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, 2024, https://doi.org/10.1049/cit2.12241.
[44] J. Zhang, S. Chen, S. Tian, W. Gong, G. Cai, and Y. Wang, "A crowd counting framework combining with crowd location," *Journal of Advanced Transportation*, vol. 2021, no. 1, p. 6664281, 2021, https://doi.org/10.1155/2021/6664281.

[45] D. Baktibayev, A. Serek, B. Berlikozha, and B. Rustauletov, "Resource-efficient sentiment classification of app reviews using a CNN-BiLSTM hybrid model," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 7, no. 3, pp. 427–433, 2025, https://doi.org/10.12928/biste.v7i3.13954.
[46] A. Shabir, K. T. Ahmed, K. Kanwal, A. Almas, S. Raza, M. Fatima, and T. Abbas, "A systematic review of attention models in natural language processing," *Statistics, Computing and Interdisciplinary Research*, vol. 6, no. 1, pp. 33–56, 2024, https://doi.org/10.52700/scir.v6i1.157.
[47] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6037–6124, 2022, https://doi.org/10.1007/s10462-022-10148-x.
[48] N. Zhang and J. Kim, "A survey on attention mechanism in NLP," in *Proc. 2023 Int. Conf. Electronics, Information, and Communication (ICEIC)*, pp. 1–4, 2023, https://doi.org/10.1109/ICEIC57457.2023.10049971.
[49] T. Wang, T. Zhang, K. Zhang, H. Wang, M. Li, and J. Lu, "Context attention fusion network for crowd counting," *Knowledge-Based Systems*, vol. 271, p. 110541, 2023, https://doi.org/10.1016/j.knosys.2023.110541.
[50] B. Tyagi, S. Nigam, and R. Singh, "A review of deep learning techniques for crowd behavior analysis," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, pp. 5427–5455, 2022, https://doi.org/10.1007/s11831-022-09772-1.
[51] A. Tomar, S. Kumar, and B. Pant, "Crowd analysis in video surveillance: A review," in *Proc. 2022 Int. Conf. Decision Aid Sciences and Applications (DASA)*, pp. 162–168, 2022, https://doi.org/10.1109/DASA54658.2022.9765008.
[52] 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, 2022, https://doi.org/10.1016/j.neucom.2021.02.103.
[53] U. Singh, J. F. Determe, F. Horlin, and P. De Doncker, "Crowd monitoring: State-of-the-art and future directions," *IETE Technical Review*, vol. 38, no. 6, pp. 578–594, 2021, https://doi.org/10.1080/02564602.2020.1803152.
[54] F. Wang, K. Liu, F. Long, N. Sang, X. Xia, and J. Sang, "Joint CNN and transformer network via weakly supervised learning for efficient crowd counting," *arXiv preprint arXiv:2203.06388*, 2022, https://doi.org/10.48550/arXiv.2203.06388.
[55] M. A. Khan, H. Menouar, and R. Hamila, "LCDnet: A Lightweight Crowd Density Estimation Model for Real-time Video Surveillance," *arXiv preprint arXiv:2302.05374*, 2023, https://doi.org/10.1007/s11554-023-01286-8.
[56] Z. Yan, P. Li, B. Wang, D. Ren, and W. Zuo, "Towards learning multi-domain crowd counting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6544–6557, 2021, https://doi.org/10.1109/TCSVT.2021.3137593.
[57] K. B. A. Hassen, J. J. Machado, and J. M. R. Tavares, "Convolutional neural networks and heuristic methods for crowd counting: A systematic review," *Sensors*, vol. 22, no. 14, p. 5286, 2022, https://doi.org/10.3390/s22145286.
[58] M. J. Babar, M. Husnain, M. M. S. Missen, A. Samad, M. Nasir, and A. K. N. Khan, "Crowd counting and density estimation using deep network—a comprehensive survey," *Authorea Preprints*, 2023, https://doi.org/10.36227/techrxiv.23256587.
[59] Y. Hao, H. Du, M. Mao, Y. Liu, and J. Fan, "A survey on regression-based crowd counting techniques," *Information Technology and Control*, vol. 52, no. 3, pp. 693–712, 2023, https://doi.org/10.5755/j01.itc.52.3.33701.
[60] 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)*, pp. 589–597, 2016, https://doi.org/10.1109/CVPR.2016.70.
[61] 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)*, pp. 1091–1100, 2018, https://doi.org/10.1109/CVPR.2018.00120.
[62] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021, https://doi.org/10.1145/3465055.
[63] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Computer Vision (ECCV)*, pp. 734–750, 2018, https://doi.org/10.1007/978-3-030-01228-1_45.
[64] Y. Li, J. Luo, X. Kong, Y. Liu, D. Fei, Y. Wang, *et al.*, "A method for regional crowd flow prediction based on crowd number estimation network," in *Proc. 2024 7th Int. Conf. Machine Learning and Natural Language Processing (MLNLP)*, pp. 1–5, 2024, https://doi.org/10.1109/MLNLP63328.2024.10800344.
[65] H. Lee and J. Lee, "TinyCount: an efficient crowd counting network for intelligent surveillance," *J. Real-Time Image Process.*, vol. 21, no. 4, p. 153, 2024, https://doi.org/10.1007/s11554-024-01531-8.

AUTHOR BIOGRAPHY

Nurmukhammed Abeuov completed his master's degree in Data Science at the Kazakh-British Technical University (KBTU) in 2022. He is currently working as a Senior ML Engineer at Biometric.Vision. His research interests include MLOps, natural language processing (NLP), and applied machine learning for computer vision. He has hands-on experience in deploying large-scale ML pipelines, optimizing model performance in production, and integrating AI solutions into real-world products.
Email: nurma.engineer@gmail.com
Orcid: https://orcid.org/0000-0001-7885-7862

Daniyar Absatov is a master's student in Software Engineering at the Kazakh-British Technical University (KBTU), Kazakhstan. He is currently working as a software engineer at Kaspi.kz. His research interests include DevOps, high-load systems, and software architecture. He has hands-on experience in building scalable backend services, integrating distributed systems, and optimizing CI/CD pipelines for production environments.
Email: da_absatov@kbtu.kz
Orcid: https://orcid.org/0009-0006-0043-6498

Yelnur Mutaliyev is a doctoral student in Software Engineering at the Kazakh National Research Technical University, Almaty. He is currently working as senior lecturer at SDU University. His research interests include Computer vision, Emotion recognition and software development. He has experience in education, QA engineering and Software testing.
Email: emutaliev11@gmail.com
Orcid: https://orcid.org/0000-0002-1755-8161

Azamat Serek is an Assistant Professor at the Kazakh-British Technical University (KBTU), Almaty, Kazakhstan. He received his Ph.D. in Computer Science from SDU University in 2024, following an M.Sc. degree in Computing Systems and Software in 2020 and a B.Sc. degree in the same field in 2018. He has published more than 15 research articles in peer-reviewed journals and conference proceedings indexed in Scopus and Web of Science and currently holds an H-index of 5 in Scopus. His research interests lie in the application of deep learning methods across multiple domains, including natural language processing, computer vision, education, as well as resource allocation and planning.
Email: a.serek@kbtu.kz
Scopus profile: https://www.scopus.com/authid/detail.uri?authorId=57207763595