FMDB *Transactions on Sustainable Computer Letters*

Enhanced Crowd Counting Using Patch-Level Annotation with YOLO Deep Learning

Saw Mya Nandar[1,*], Hlaing Htake Khaung Tin[2], Bhopendra Singh[3], Mykhailo Paslavskyi[4], Andino Maseleno[5], András Szeberényi[6]

[1,2]Faculty of Computer Systems and Technologies, University of Computer Studies, Yangon, Yangon Region, Myanmar.
[3]Department of Engineering, Amity University Dubai, Dubai, United Arab Emirates.
[4]Department of Computer Science, Ukrainian National Forestry University, Lviv, Lviv Oblast, Ukraine.
[5]Department of Information Systems, Institut Bakti Nusantara, Pringsewu, Lampung, Indonesia.
[6]Institute of Marketing and Communication, Budapest Metropolitan University, Budapest, Hungary.
sawmyananda2025@gmail.com[1], hlainghtakekhaungtin@gmail.com[2], bhopendrasingh@yahoo.com[3], mykhailo.paslavskyi@nltu.edu.ua[4], andino.maseleno@ibnus.ac.id[5], aszeberenyi@metropolitan.hu[6]

Abstract: Accurate crowd counting is of great importance to a wide range of real-world applications such as public safety monitoring, traffic management, event organisation, urban planning and emergency response. However, existing crowd counting techniques often suffer from severe occlusions, large variations in crowd density, perspective distortion and complex background environments, leading to reduced detection accuracy. In this paper, a new crowd counting framework is proposed based on patch-level annotation and YOLO (You Only Look Once) deep learning architecture to improve detection performance and real-time processing capability. The suggested method separates high-resolution crowd photos into smaller patches. It performs accurate patch-level annotation, allowing the model to learn local crowd characteristics, minimise annotation complexity, and better localise tightly packed people. Processing these image patches with YOLO's quick object detection allows reliable crowd estimation with low inference time. The proposed framework surpasses state-of-the-art approaches in accuracy, robustness, and computational efficiency, as demonstrated by extensive trials on benchmark crowd-counting datasets. A performance study using standard measures demonstrates that identification and counting accuracy improve significantly across diverse crowd densities and challenging conditions. The suggested system is effective and scalable for intelligent surveillance, smart city monitoring, public event management, and real-time crowd analysis.

Keywords: Crowd Counting; Patch-Level Annotation; Deep Learning (DL); Event Management; Real-Time Surveillance; Computer Vision; Intelligent Crowd Analysis; Smart City Monitoring.

Received on: 19/07/2025, **Revised on:** 12/10/2025, **Accepted on:** 23/10/2025, **Published on:** 09/08/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2026.000751

Cite as: S. M. Nandar, H. H. K. Tin, B. Singh, M. Paslavskyi, A. Maseleno, and A. Szeberényi, "Enhanced Crowd Counting Using Patch-Level Annotation with YOLO Deep Learning," *FMDB Transactions on Sustainable Computer Letters*, vol. 4, no. 3, pp. 169–177, 2026.

1. Introduction

[*]Corresponding author.

Crowd counting is one of the most important research fields in computer vision due to its wide range of applications, including public safety, intelligent transportation systems, urban planning, surveillance, smart cities, event management, and disaster response. The rapid growth of cities and the population often results in densely crowded public spaces, such as transportation terminals, shopping malls, railway stations, airports, sports arenas, religious gatherings, educational institutions, and entertainment venues [2]. Manual monitoring of these crowded environments has become increasingly difficult, time-consuming, and error-prone. Hence, intelligent automated crowd-counting systems have become crucial for real-time, accurate estimation of the number of individuals to support informed decision-making and public safety [3]. Accurate crowd estimation is useful for a variety of government and commercial applications. In large public gatherings, authorities need to continuously monitor crowd density to prevent dangerous overcrowding that can lead to stampedes, congestion or failure of emergency evacuation [7]. Real-time crowd analysis also allows traffic management authorities to optimise pedestrian movement, improve transportation planning, and allocate security personnel more efficiently. Crowd counting information is useful for infrastructure planning, resource allocation and environmental monitoring in smart city environments. Crowd analytics allows retail businesses to gain insights into customer behaviour, optimise store layouts, and enhance marketing strategies.

Moreover, health care institutions and emergency responders are deploying crowd monitoring systems to enforce social distancing during disease outbreaks or to coordinate rescue missions during natural disasters [10]. Crowd counting is a very challenging computer vision problem, although it is widely applicable. In real-world images, there are often thousands of people with large variations in appearance, pose, illumination and viewing angle. People who are closer to the camera are normally larger than those farther away due to perspective distortion. Severe occlusions occur when people partially or fully cover each other in crowded scenes, making it extremely hard to detect individual people. Furthermore, the counting task is further complicated by dynamic lighting conditions, shadows, background clutter, weather changes, camera motion and varying image resolutions. These challenges severely impact the performance of traditional image processing techniques and demand more sophisticated, flexible intelligent solutions that can adapt to complex visual environments [4]. Previous approaches to crowd counting relied heavily on handcrafted image features and classical machine learning algorithms. These methods extracted low-level visual descriptors, such as edges, corners, textures, gradients, foreground segmentation, optical flow, Haar features, Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT). Afterwards, statistical regression models were trained to estimate the total number of people using these handcrafted features [11].

These techniques performed reasonably well in low-density environments with limited occlusion, but they degraded significantly in highly congested scenes. The handcrafted features relied heavily on manually designed feature representations. They were sensitive to changes in illumination, viewing angles, crowd density, and background complexity, so they were unable to generalise across different datasets [13]. Later work proposed density-mapping methods to address some of these limitations. Density-based algorithms produce continuous density maps rather than directly estimating individual detections, and the estimated crowd count is the integral of the density values [14]. For images with dense populations, density estimation significantly improved counting accuracy because it did not rely on explicit person detection. Several density estimation models based on regression and convolutional neural networks (CNN) have shown promising performance on benchmark datasets. However, density map methods typically result in blurred representations without precise localisation of individuals. Furthermore, they require computationally intensive processing and a large number of pixel-level annotations, which makes them less suitable for real-time deployment in surveillance systems where rapid decision-making is essential [12]. The past decade has seen rapid development of deep learning, which has profoundly changed computer vision research. Deep neural networks learn hierarchical feature representations directly from raw image data, eliminating the need for handcrafted feature engineering [16].

Convolutional Neural Networks (CNNs) have been very successful on many vision tasks, including image classification, semantic segmentation, object detection, facial recognition, medical imaging and autonomous driving. These networks can model complex spatial patterns and high-level semantics, leading to significantly improved accuracy and robustness over traditional methods [15]. Object detection is one of the most influential fields of deep learning, which allows automatic recognition and localisation of multiple objects in a single image. Modern object detectors can be divided into two-stage and single-stage detection frameworks. Two-stage detectors like Faster R-CNN, Mask R-CNN, and Cascade R-CNN first propose candidate object regions and then classify and localise the object in each region. These techniques have high detection accuracy. Nevertheless, their real-time implementations are often hindered by their computational complexity. Instead, single-stage detectors perform localisation and classification in an integrated manner, which gives a significant speed-up with competitive detection accuracy [17]. For single-stage detectors, the YOLO (You Only Look Once) algorithm has gained significant popularity for its good trade-off between detection speed and accuracy. YOLO frames object detection as a single regression problem, enabling full-image interpretation in a single forward pass of the neural network. Unlike region proposal methods, YOLO simultaneously predicts object locations, class probabilities and confidence scores, resulting in substantially faster inference [18]. YOLO has been released in several versions, each improving detection accuracy, computational efficiency, robustness and scalability.

Currently, YOLO is widely used in autonomous vehicles, intelligent surveillance, industrial automation, robotics, and agriculture, health care, and security systems for real-time object detection [19]. Although YOLO has achieved excellent performance in general object detection, applying it directly to crowd counting in extremely dense environments remains a challenge. The images contain thousands of people, but they are very small, heavily overlapped, and partially obscured by severe occlusions. The limited spatial resolution assigned to each object complicates accurate localisation. Therefore, CT processing at the image level usually cannot identify persons or combine persons who are close to each other into a single detection, thereby reducing counting accuracy. Recently, image partition strategies have attracted significant interest in dense object detection applications to overcome these challenges. Rather than treating the image as a single input, it is broken into multiple smaller regions, or patches [1]. In each patch, fewer individuals are available, allowing the neural network to focus on local spatial information at higher resolution. Patch-based processing substantially reduces object overlap, preserves fine-grained visual details, and allows for more accurate localisation of densely packed people [3]. Additionally, smaller patches make annotation easier, as annotators label only local areas rather than full high-density scenes, reducing labelling complexity and enhancing annotation consistency.

The patch-level annotation has several important advantages over conventional image-level annotation: First, it enables the model to learn localised crowd distribution patterns more efficiently by focusing on small regions with relatively uniform density characteristics [8]. Second, it reduces annotation ambiguity by overlapping people in crowded crowds. Third, it enables efficient parallel processing during both training and inference, thereby improving computational performance [5]. Fourth, patch-level learning can enhance generalisation across various crowd densities because the model learns representative local features rather than relying solely on global image characteristics. These benefits combined lead to improved detection accuracy, particularly in heavily cluttered scenes where typical object detectors generally perform poorly. Patch-level annotation integrated with the YOLO architecture is a promising direction for intelligent crowd-counting systems. Combining YOLO's localisation capability with localised patch-based learning enables accurate counting with real-time performance [7]. The detection network operates independently on each image patch, enabling detailed feature extraction and precise localisation of individuals, even in dense crowd areas [2]. After all the patches have been processed, the local counts are combined to obtain the final crowd count for the whole image. The strategy balances computational efficiency and detection accuracy effectively and is compatible with the existing real-time surveillance infrastructure. This paper proposes a framework based on this patch-based philosophy to improve crowd counting performance under challenging real-world conditions.

First, the high-resolution crowd images are segmented into small patches by a regular grid-based segmentation strategy. Each patch is meticulously annotated to improve the accuracy of local crowd distributions [4]. Then, the annotated patches are used to train a YOLO-based detection network that can accurately recognise people in different crowd densities, views and illumination conditions. During inference, the trained model analyses each image patch independently and combines the local predictions to provide a full crowd count [1]. Such a modular processing strategy can greatly enhance detection accuracy while maintaining fast inference speed for practical deployment. Researchers evaluate the performance of the proposed framework on publicly available benchmark datasets for crowd counting that span a wide range of environmental conditions and crowd densities [6]. The performance of the proposed approach is evaluated using standard performance metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Precision, Recall, F1-score, and inference time. Patch-level annotation significantly improves YOLO performance for detecting densely packed individuals and reduces counting errors compared to image-level processing, according to the experimental results [8]. The framework can provide accurate estimates of crowd counts in sparse, medium-density, and highly dense scenes while maintaining real-time computational performance [9]. Besides surveillance, the proposed crowd-counting framework can be widely applied across many intelligent systems. The crowd information generated can be exploited by smart city managers for adaptive traffic control, pedestrian flow optimisation, emergency evacuation planning, and infrastructure management.

Transport authorities might monitor passenger numbers at railway stations, airports and bus terminals to improve operational efficiency and passenger safety [1]. Event organisers can control venue capacity and allocate security personnel adaptively based on real-time crowd analysis. Retail businesses can use crowd analytics to improve customer service and store management [11]. In addition, precise crowd estimations can be utilised by healthcare agencies and disaster management organisations to support emergency response, pandemic control measures and humanitarian relief operations. To sum up, crowd counting remains a very challenging problem in modern computer vision due to dense occlusions, perspective distortion, background clutter, and large variations in crowd density [2]. While deep learning has greatly improved counting performance, further improvements are required to achieve reliable, real-time operation in highly congested scenes [13]. The proposed combination of patch-level annotation and YOLO deep learning architecture aims to tackle these issues by integrating localised feature learning and efficient object detection. The resulting framework can improve counting accuracy, reduce annotation complexity and maintain high computational efficiency. It makes the framework suitable for intelligent surveillance, public safety monitoring, smart city management and other real-time crowd analysis applications. This research is a step towards scalable, robust and practical crowd counting systems that can meet the growing demands of modern intelligent urban environments [15].

2. Related Work

Khan et al. [1] provided a comprehensive review of crowd monitoring and localisation using deep convolutional neural networks. The authors reviewed the evolution of crowd counting from traditional image processing techniques to state-of-the-art deep learning frameworks. The study emphasises that convolutional neural networks greatly enhance feature extraction, localisation accuracy, and crowd estimation in challenging situations including occlusions, scale variations, and complex backgrounds. They found that deep learning has emerged as the dominant paradigm for intelligent crowd-monitoring systems. Chavan and Purohit [2] reviewed machine learning techniques for crowd counting and examined the transition from handcrafted feature-based methods to deep learning-based methods. They point out in their review that traditional algorithms based on Histogram of Oriented Gradients (HOG), Support Vector Machines (SVM) and regression techniques have poor generalisation in dense crowd environments [12]. The authors conclude that YOLO-based object detection and CNN-based density estimation are promising approaches to achieve high accuracy and real-time performance simultaneously. Wang et al. [3] proposed a weakly supervised crowd counting framework based on Convolutional Neural Networks (CNNs) and transformers. The model integrates local feature extraction and global context learning to enhance crowd counting and reduce the need for annotations. Researchers present experiments demonstrating that incorporating transformer networks improves counting accuracy in very crowded scenes, where traditional CNN models tend to fail. Elsepae et al. [4] studied the state of the art in deep learning-based crowd counting in complex environments. In their review, they considered challenges such as occlusions, perspective distortion, illumination, background clutter and crowdedness.

The authors reviewed emerging research trends, including lightweight neural networks, attention mechanisms, transformer-based architectures, self-supervised learning, and hybrid detection models, to improve computational efficiency and counting accuracy. Sutrisno et al. [5] optimised the YOLO11 object detection framework for dense crowd counting by introducing head-detection fine-tuning. The proposed method localises heads rather than full-body detection, enabling better performance under severe occlusion conditions. Their experiments showed improved counting accuracy while maintaining the YOLO architecture's real-time detection capability. A people counting system based on YOLO and clustering algorithms for autonomous mobile robots was proposed by Gomulka et al. [6]. They use YOLO to detect pedestrians and estimate crowd size using spatial clustering while the robot is navigating. The proposed approach achieved reliable counting performance in dynamic environments, while maintaining computational efficiency suitable for real-time robotic applications. Wang et al. [7] proposed a patch-based crowd-counting method that combines Convolutional Neural Networks with density map estimation. They divide high-resolution images into small patches so the model can learn localised crowd characteristics more effectively. The experimental results showed that counting accuracy improved in highly congested scenes by alleviating occlusion effects while retaining local spatial information. Deng et al. [8] provided a comprehensive survey of deep learning techniques for crowd counting. The existing approaches are categorised into detection-based, regression-based, density estimation, attention-based, transformer-based, and hybrid methods [1].

Reviewers concluded that combining localised feature learning with efficient object detection architectures is one of the most promising research directions for future intelligent crowd-counting systems. Li et al. [9] proposed a dilated convolutional neural network-based crowd counting framework, CSR-Net. The model exploits dilated convolutions to enlarge the receptive field without sacrificing feature resolution, resulting in accurate density estimation in extremely crowded scenes [3]. Their experimental results indicated that CSR-Net is a benchmark architecture for density-map-based crowd counting, achieving high counting accuracy and robustness. The existing literature shows that crowd counting has evolved from handcrafted feature-based methods to sophisticated deep learning frameworks. Density estimation methods produce accurate counts but require heavy pixel-wise annotation and high computational costs [1]. Approaches for object detection, particularly those based on the YOLO architecture, can operate in real time but may suffer from reduced accuracy in very dense scenes with severe occlusions. Patch-based learning strategies improve localisation by focusing on local crowd areas, while transformer-based models provide better contextual understanding of the whole image [8]. These results suggest that a combination of patch-level annotation and an improved YOLO detection framework can reach an efficient trade-off between annotation efficiency, computational speed and counting accuracy. Motivated by these findings, this study proposes a modified patch-wise labelling strategy in conjunction with a modern YOLO network to improve crowd counting performance in dense and difficult real-world surveillance scenarios [2].

3. Proposed Methodology

3.1. Patch-Wise Labelling Method

In dense scenes with dense crowds, annotating entire images is cumbersome and prone to error. To address this, the system proposed here dissects input images into small, fixed-size patches. Each patch is annotated individually, allowing the model to focus on local crowd patterns and reduce annotation uncertainty. Patch-wise labelling simplifies annotation and helps the network learn rich spatial information, especially in dense regions.

3.2. YOLO Network Adaptation

The YOLO network, with its speed and accuracy in object detection, is used to detect people in every image patch. The model is fine-tuned to focus on detecting small patches containing people by training on patch-level labelled datasets. Localised detection enhances the ability to distinguish and occlude people, a common problem in crowd counting. At inference, the system runs the trained YOLO network on each patch, counting the number of detected people. The overall crowd count is obtained by summing the counts across all patches. This type of module-based method provides parallel processing and real-time capability.

3.3. Training and Implementation Details

The network is initialised using publicly available crowd datasets, with images preprocessed into patches. Data augmentation, including rotation, scaling, and brightness adjustments, is applied to improve generalisation. The YOLO model is fine-tuned for optimal detection on patch-level inputs, balancing precision and recall. Figure 1 illustrates the general workflow of the proposed patch-level crowd-counting framework using the YOLO object detection model. The framework first decomposes a large crowd image into smaller image patches and then performs object detection to improve crowd-counting accuracy. The Full Image is the original image input from a surveillance camera or monitoring system, and the process begins with it. Crowd images often exhibit variations in density, scale and occlusions, so processing the entire image directly may reduce detection accuracy. The next step is Image Partitioning into Patches, in which the original image is split into several smaller, equally sized patches. This division allows the detection model to focus on local regions rather than the entire image, reducing the impact of overlapping people and improving feature extraction in dense areas. Patch-wise Labelling is performed after partitioning. Each image patch is annotated separately to indicate the locations of people within that patch.

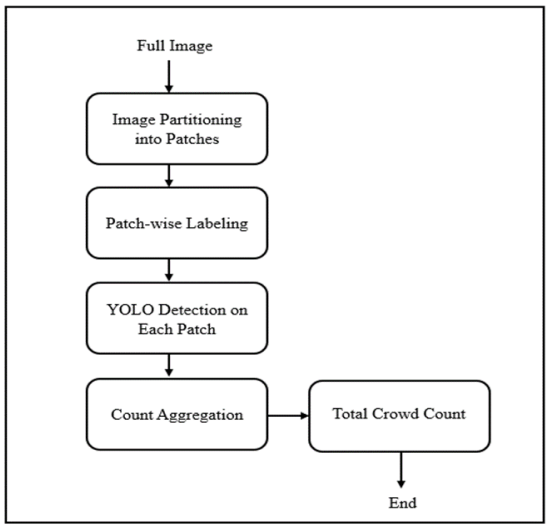

Figure 1: Overview of the proposed crowd counting method

Localised annotation provides more accurate ground-truth information during model training. It enables the detector to learn the fine-grained characteristics of the crowd more effectively than whole-image annotations. YOLO Detection on Each Patch stage works with the labelled patches. YOLO (You Only Look Once) produces bounding boxes and confidence scores and independently detects people in each patch. YOLO can accurately predict close or partially visible objects by processing smaller image regions while still running in real time. The outputs of all patches are combined in the Count Aggregation stage after detection. Researchers sum the counts for each patch, after removing duplicate detections at patch boundaries, to obtain a consistent estimate of the total number of people in the original image. Finally, the aggregated result is given as the Total Crowd Count, indicating the estimated number of persons existing in the full scene. The results can be used for surveillance, public safety monitoring, event management, transportation systems and smart city applications. The process ends. In summary, the proposed framework integrates patch-level image partitioning and YOLO-based object detection to achieve higher counting accuracy, improved robustness in dense crowd scenarios, lower annotation complexity and efficient real-time performance compared with traditional whole-image crowd counting methods. To formalise the patch-wise counting process:

$$C = \sum_{(i=1)}^{N} c_i \quad (1)$$

Where:

- C = Total estimated crowd count in the full image
- N = number of patches the image is divided into
- ci = number of people detected in the ith patch by the YOLO network

$$c_i = \sum_{(j=1)}^{(M_i)} 1\left(s_{(ij)} \geq \tau\right) \quad (2)$$

Where:

- Mi = number of bounding boxes detected in patch i
- sij = Confidence score of the jth detection in patch i
- 1(·) = Indicator function (1 if true, 0 otherwise)

4. Experimental Results and Evaluation

The proposed crowd counting system is evaluated on benchmark datasets such as ShanghaiTech Part A and UCF_CC_50, which are well-known for their diverse crowd densities and complicated scenes. Counting accuracy is evaluated using performance metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

Table 1: Performance comparison of benchmark datasets

Method	Dataset	MAE	RMSE	Inference Time (ms)
MCNN [15]	Shanghai Tech A	110.2	173.2	120
CSRNet [9]	Shanghai Tech A	68.2	115.0	150
YOLO-based (Baseline)	Shanghai Tech A	75.5	130.1	45
Proposed Patch-wise YOLO	Shanghai Tech A	62.8	108.4	50
MCNN	UCF_CC_50	295.8	320.9	130
CSRNet	UCF_CC_50	266.1	397.5	160
YOLO-based (Baseline)	UCF_CC_50	280.4	350.2	55
Proposed Patch-wise YOLO	UCF_CC_50	245.7	320.1	60

Table 1 summarises the evaluation on benchmark datasets (i.e., ShanghaiTech Part A and UCF_CC_50). Table 1 provides a comparative evaluation of different crowd counting methods across two benchmark datasets: ShanghaiTech Part A and UCF_CC_50. The performance measures considered here are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and inference time in milliseconds (ms), where lower values indicate better performance.

4.1. Shanghai Tech Part-A Dataset

MCNN is a traditional multi-column CNN baseline, achieving an MAE of 110.2 and RMSE of 173.2 at the cost of an inference time of 120 ms (Figure 2).

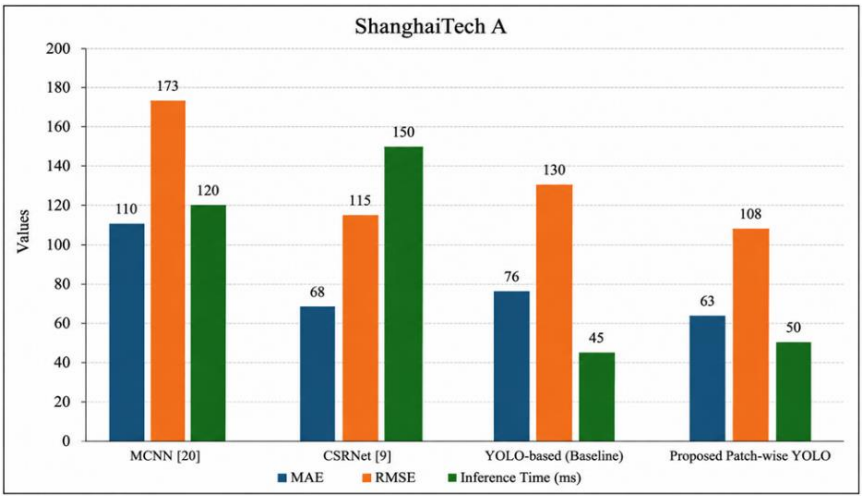

Figure 2: Performance comparison Shanghai Tech A

Even though it laid the foundation for crowd counting, its performance shows shortcomings in handling scale changes and densely packed crowds. CSRNet significantly improves accuracy, reducing MAE to 68.2 and RMSE to 115.0. But all this comes at the cost of increased inference time (150 ms), indicating greater computational complexity from deeper, and more dilated convolutional layers. The YOLO-based baseline, focused on object detection rather than density estimation, achieves 45 ms of inference time, making it suitable for real-time applications. However, it has relatively higher MAE (75.5) and RMSE (130.1) than CSRNet, indicating a trade-off between speed and accuracy. The proposed Patch-wise YOLO method outperforms all the methods compared in terms of accuracy, with an MAE of 62.8 and RMSE of 108.4, while still maintaining competitive inference time (50 ms). This shows that dividing the image into patches and using the YOLO network maximises detection accuracy and trustworthiness, particularly in densely populated regions, while incurring minimal speed loss.

4.2. UCF_CC_50 Dataset

This dataset has a more challenging environment due to extremely high crowd density and dynamic scenes. For this dataset, the MCNN baseline again obtains an MAE of 295.8 and an RMSE of 320.9, due to the difficulty in handling such complex data. CSRNet improves on these to an MAE of 266.1 and an RMSE of 397.5 but at an increased inference time of 160 ms. The YOLO-based baseline provides faster inference (55 ms) but relatively larger MAE (280.4) and RMSE (350.2), indicating that direct detection may not be effective in high-crowd situations. The proposed Patch-wise YOLO method is the most accurate on this dataset, achieving 245.7 MAE and 320.1 RMSE with a mere 60 ms inference time (Figure 3).

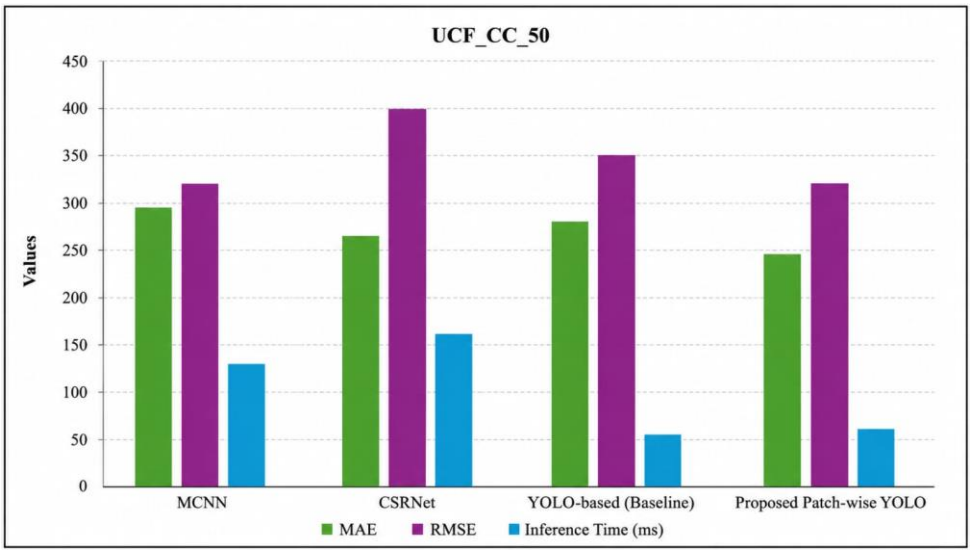

Figure 3: Performance comparison UCF_CC_50

This ensures the effectiveness of patch-wise labelling in boosting detection performance for extremely dense scenes, balancing accuracy and efficiency well. The proposed patch-wise YOLO approach demonstrates significant improvements over both baseline detection-based and typical density-based models in terms of accuracy (lower MAE and RMSE) while maintaining reasonable inference speeds for real-world applications. This supports the main hypothesis that local patch-level inspection, combined with a powerful detection network, can more effectively address the difficulties inherent in dense, complex crowd-counting cases. Results show that patch-wise annotation, combined with YOLO detection, significantly improves counting accuracy compared to baseline methods. The system shows improved occlusion handling and dense crowd management, with lower MAE and RMSE values. In addition, real-time inference speed is maintained, which enhances the method's suitability for real-world applications.

5. Conclusion

This paper presents an improved crowd-counting approach that uses patch-level tagging with the YOLO detection network. By partitioning large images into small, manageable patches and using localised, precise labelling, the approach presented here dramatically improves the model's ability to detect people in crowded, complex crowd scenes. The patch-wise approach enables the model to focus on local spatial information, thereby significantly improving detection accuracy, especially in heavily occluded and unevenly distributed crowd scenarios. Integration with the YOLO detection network provides an optimal trade-off between accuracy and computational cost. Fast inference owing to YOLO's one-stage architecture provides consistency and makes the proposed approach ideal for real-time surveillance applications and crowd management systems. The empirical

results on widely used benchmark datasets, such as the ShanghaiTech and UCF_CC_50 datasets, indicate that the proposed method outperforms existing baseline models in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and inference speed.

5.1. Future Work

Despite these promising results, several potential directions for further work remain. Firstly, the model developed here is primarily for static images; therefore, extending the methodology to video sequences could offer dynamic crowd tracking and analysis. Temporal information from recurrent neural networks (RNNs) or 3D convolutional networks could further stabilise counting and reduce flickering in per-frame counts. Secondly, recent advances in transformer-based architectures have shown tremendous potential across a range of vision tasks by handling long-range relationships and global context. Future work could apply these architectures to crowd counting to further enhance accuracy, particularly in highly crowded and complex scenes. Thirdly, weakly supervised and semi-supervised learning techniques could be integrated to counteract the manpower-intensive nature of annotation in dense crowds. By reducing reliance on complete patch-wise annotation, the techniques would improve annotation efficiency while maintaining or even improving performance. This paper advances the state of the art in crowd counting by incorporating patch-level analysis into a computationally efficient detection network. The proposed method provides a solid basis for real-time crowd monitoring, and further investigation will focus on broadening its applicability, resilience, and annotation effectiveness.

Acknowledgement: The authors sincerely acknowledge the valuable academic support, research facilities, and institutional encouragement provided by the University of Computer Studies, Amity University Dubai, Ukrainian National Forestry University, Institut Bakti Nusantara, and Budapest Metropolitan University, which contributed to the successful completion of this research.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding Statement: This research did not receive any funding.

Conflicts of Interest Statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics and Consent Statement: The study was conducted in accordance with ethical guidelines. Participants were assured of the confidentiality and anonymity of their responses.

References

1. A. Khan, J. A. Shah, K. Kadir, W. Albattah, and F. Khan, “Crowd monitoring and localization using deep convolutional neural network: A review,” *Appl. Sci.*, vol. 10, no. 14, pp. 1–17, 2020.
2. D. Chavan and A. Purohit, “Machine Learning Techniques for Crowd Counting: A Survey,” *Int. J. Comput. Appl.*, vol. 185, no. 37, pp. 1–8, 2023.
3. F. Wang, K. Liu, N. Wei, N. Sang, X. Xia, and J. Sang, “Weakly supervised crowd counting with joint CNN and transformer network,” *Pattern Recognit.*, vol. 171, no. 3, p. 112077, 2026.
4. H. F. Elsepae, H. M. El-Hoseny, E. K. I. Hamad, and E. S. M. El-Rabaie, “Deep learning for crowd counting in complex environments: Challenges and novel trends,” *Discov. Comput.*, vol. 29, no. 2, pp. 1–34, 2026.
5. J. Sutrisno, S. Winarno, and A. Affandy, “Optimizing YOLO11 for dense crowd counting under severe occlusion via head-detection fine-tuning,” *J. Tech. Inform. (JUTIF)*, vol. 7, no. 2, pp. 1871–1885, 2026.
6. K. Gomulka, P. Wozniak, and T. Krzeszowski, “People counting using YOLO-based detection and clustering for a mobile robot,” *Sensors*, vol. 26, no. 15, pp. 1–25, 2026.
7. Y. Wang, W. Zhang, Y. Liu, and J. Zhu, “Multi-density map fusion network for crowd counting,” *Neurocomputing*, vol. 397, no. 7, pp. 31–38, 2020.
8. L. Deng, Q. Zhou, S. Wang, J. M. Górriz, and Y. Zhang, “Deep learning in crowd counting: A survey,” *CAAI Trans. Intell. Technol.*, vol. 9, no. 5, pp. 1043–1077, 2023.
9. Y. Li, X. Zhang, and D. Chen, “CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, Utah, United States of America, 2018.
10. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in Computer Vision – ECCV 2016, *Springer*, Cham, Switzerland, 2016.

11. R. Nasir, Z. Jalil, M. Nasir, T. Alsubait, M. Ashraf, and S. Saleem, "An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8," *Artif. Intell. Rev.*, vol. 58, no. 4, pp. 1–50, 2025.
12. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint*, 2018. [Accessed by 12/05/2025].
13. Y. Lei, Y. Liu, P. Zhang, and L. Liu, "Towards using count-level weak supervision for crowd counting," *Pattern Recognition*, vol. 109, no. 1, p. 107616, 2021.
14. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, Nevada, United States of America, 2016.
15. S. Zhang, Z. Wang, Q. Liu, F. Wang, W. Ke, and T. Zhang, "Crowd counting with sparse annotation," *Pattern Recognit.*, vol. 180, no. 12, p. 114000, 2026.
16. T. Zhao, Z. Shen, D. Li, P. Zhong, and J. Tan, "A Scale-Aware local Context aggregation network for Multi-Domain shrimp counting," *Expert Systems with Applications*, vol. 267, no. 4, p. 126179, 2025.
17. W. Mansouri, M. A. Alohali, H. Alqahtani, N. Alruwais, M. Alshammeri, and A. Mahmud, "Deep convolutional neural network-based enhanced crowd density monitoring for intelligent urban planning on smart cities," *Sci. Rep.*, vol. 15, no. 2, pp. 1–24, 2025.
18. Z. Huang, R. Sinnott, and Q. Ke, "Crowd counting using deep learning in edge devices," in *Proc. IEEE/ACM 8th Int. Conf. Big Data Comput., Appl. Technol. (BDCAT)*, Leicester, United Kingdom, 2021.
19. 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. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, Nevada, United States of America, 2016.