engineering proceedings

MDPI

Proceeding Paper

ClassRoom-Crowd: A Comprehensive Dataset for Classroom Crowd Counting and Cross-Domain Baseline Analysis †

Wenqian Jiang [1], Xiaohua Huang [1,2,3,*], Qun Zhao [4] and Sheng Liu [4]

[1] Oulu School, Nanjing Institute of Technology, Nanjing 21167, China; y00450220206@njit.edu.cn
[2] The Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Southeast University, Nanjing 21189, China
[3] Research Center for Learning Science, Southeast University, Nanjing 21189, China
[4] School of Computer Engineering, Nanjing Institute of Technology, Nanjing 21167, China; qunzhaoqz@outlook.com (Q.Z.); lius@njit.edu.cn (S.L.)
* Correspondence: xiaohua.huang@njit.edu.cn
† Presented at the 1st International Conference on AI Sensors & the 10th International Symposium on Sensor Science, Singapore, 1–4 August 2024.

Abstract: In recent years, with the rise of smart education, crowd counting technology has garnered increasing attention for its applications in educational environments. During epidemic outbreaks, ensuring the accurate monitoring of student density in classrooms and other public educational spaces has become crucial. Additionally, in day-to-day educational management, tasks such as optimizing the allocation of teaching resources and planning educational spaces heavily depend on precise crowd counting techniques. Current methods predominantly employ convolutional neural networks, which require large datasets for training. However, there is a lack of specialized datasets tailored to specific educational scenarios and extreme conditions. Furthermore, to enhance crowd movement predictions in real-world applications, these methods must demonstrate temporal coherence. To address these gaps, this paper introduces the ClassRoom-Crowd dataset, specifically designed for crowd counting in educational settings with extreme conditions and temporal continuity. The dataset consists of 7571 images with 172,898 annotated objects. Additionally, baseline results using state-of-the-art crowd counting algorithms under a cross-condition protocol are provided. The experimental findings reveal that domain shifts significantly impact model performance, underscoring the need for domain adaptation methods in crowd counting research, particularly within the smart education context.

Keywords: crowd counting; crowd analysis; density map estimate

check for updates

Academic Editor: Po-Liang Liu

Published: 11 February 2025

Citation: Jiang, W.; Huang, X.; Zhao, Q.; Liu, S. ClassRoom-Crowd: A Comprehensive Dataset for Classroom Crowd Counting and Cross-Domain Baseline Analysis. *Eng. Proc.* **2024**, *78*, 10. https://doi.org/10.3390/engproc2024078010

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Crowd counting is a vital task not only in video surveillance and public safety but also within the domain of smart education. Accurate crowd counting and localization are essential for managing classroom occupancy, optimizing the use of educational spaces, and ensuring student safety. In educational environments, tasks such as monitoring student flow and managing crowd density during school events depend heavily on reliable crowd counting techniques [1,2]. Moreover, during epidemic outbreaks and the subsequent return to in-person learning, real-time crowd counting has become increasingly important for ensuring safe distancing and managing student movements in educational facilities.

Recent advancements in deep learning have led to significant improvements in crowd counting methodologies, with cutting-edge models utilizing deep neural networks to estimate crowd density [3,4]. However, these models often require large amounts of

annotated data for training. While synthetic data can be used to mitigate the need for real-world data, domain shifts between synthetic and real-world environments frequently result in poor generalization. As a result, the use of real-scene images is critical for improving model performance [5]. Unfortunately, existing publicly available datasets lack sufficient diversity, particularly for extreme environments and classroom environments, limiting their effectiveness in more challenging contexts.

Most available datasets consist of images from various internet scenes, with limited coverage of extreme conditions, such as low lighting or heavy occlusion. This leads to poor cross-scene transferability and reduced robustness in crowd counting models. To address these limitations, we introduce the ClassRoom-Crowd dataset, specifically designed for crowd counting in classroom settings under extreme conditions. The ClassRoom-Crowd dataset offers several key advantages. (1) Scale and specificity. It is a dataset dedicated to classroom environments, comprising 7571 images. Of these, 7258 were captured under normal lighting, while 313 were taken in low-light conditions, with a total of 172,898 annotated targets. (2) Diverse conditions. The dataset includes images with uneven lighting and low-light environments, improving the cross-scene transferability of models. (3) Occlusion challenges. Each image contains dense occlusions, with students engaged in various activities, leading to highly overlapping targets, thus providing rich data for researching crowd counting under occlusion challenges.

To evaluate the performance of crowd counting models on this dataset, we conducted experiments using existing mainstream and traditional methods. Our analysis yielded the following insights: (1) Crowd counting in extreme environments remains a significant challenge for current algorithms; (2) Existing models struggle to handle both sparse and dense crowd counting tasks. Models that perform well in large-scale crowd counting often exhibit suboptimal performance in sparse crowd counting scenarios.

The main contributions are as follows:

1. We present the ClassRoom-Crowd dataset, comprising 7571 images with a total of 172,898 annotated targets, covering both normal and low-illumination environments.
2. We provide baseline results using the state-of-the-art models on this new dataset. The proposed dataset serves as a valuable resource for evaluating and benchmarking crowd counting models in specific real-world tasks, promoting future research into crowd counting problems in challenging scenarios.

The remainder of this paper is structured as follows: Section 2 reviews related work on crowd counting and datasets. Section 3 details the creation of the ClassRoom-Crowd dataset. Section 4 presents experimental results on the new database and discusses the performance of different models. Finally, Section 5 offers concluding remarks.

2. Related Work

2.1. Crowd Counting Methods

Crowd counting methods across different scenarios are typically classified into three categories: detection-based methods, regression-based methods, and density estimation-based methods. The detection-based category primarily follows a detect-then-count strategy, employing sliding window detectors to identify human bodies [6–8]. These methods require classifiers trained to extract features from full-body representations. However, 2D image-based methods face significant challenges in dealing with occlusions, variable lighting conditions, and densely populated scenes. This limitation has sparked increased interest in depth-based techniques, which provide additional spatial information. Depth data can enhance the accuracy of crowd counting by enabling more precise localization, especially in scenarios where occlusion is prominent. For instance, Khan et al.'s study [9] demonstrated that depth images could successfully detect multiple humans in challeng-

ing environments, underscoring the benefits of integrating depth information into crowd analysis models.

Regression-based methods were developed to overcome the shortcomings of detect-then-count approaches in crowded scenes. These methods directly predict crowd density from feature vectors [10–12]. Early regression-based methods relied on manually crafted features, such as SIFT and LBP, to train regression models for crowd estimation [13]. For example, Chan et al. [14] utilized a Gaussian process regression model to extract features like edges and textures for crowd counting. More recent developments have shifted toward end-to-end learning using deep CNNs. For example, Wang et al. [15] adapted the AlexNet architecture by replacing its final fully connected layer with a single neuron to estimate crowd counting. While regression-based methods generally outperform detection-based approaches, they often fail to fully exploit the available point-level supervision from annotations.

Density estimation-based methods in crowd counting involve deriving the count by summing the estimated density map of the scene. Initially proposed by Lempitsky and Zisserman [16], this technique transforms point annotations into density maps using Gaussian kernels as "ground truth" and trains the model via least squares optimization. Modern implementations utilize the feature extraction capabilities of deep neural networks (DNNs), leading to significant performance improvements [17]. However, because these methods rely on a pixel-to-pixel supervised loss function, their effectiveness is highly dependent on the accuracy of the generated ground truth density maps, which can be a limiting factor in achieving robust results.

2.2. Generating Crowd Density Map Methods

In modern crowd counting techniques, network models are trained on density maps designed to approximate crowd counts through pseudo-density maps embedded in images. These maps are generated by transforming the annotated central coordinates of individuals' heads in crowd images into density representations using Gaussian kernels. The ability of models to effectively learn features from these maps is highly dependent on the quality of the generated density maps. The creation of these maps is influenced by the configuration of the Gaussian kernels, which is typically categorized into three main approaches: geometric-adaptive, fixed Gaussian kernel, and content-aware annotation techniques. The geometric-adaptive method dynamically adjusts the parameters of the density map by leveraging geometric features of crowd distribution, allowing for improved accuracy and adaptability in complex and varied crowd scenes [18]. Fixed Gaussian kernel methods apply a fixed Gaussian kernel at each person's location in the image, with all kernels superimposed to generate an overall density map. While this method is computationally efficient and easy to implement, it tends to be less accurate in handling complex and varied crowd scenes with high variability [19]. The content-aware annotation method uses object detection and semantic segmentation to intelligently annotate images based on their content, making it well suited to complex scenes. However, this approach requires extensive datasets and significant computational resources to achieve optimal performance [20].

2.3. Database for Crowd Count

Recent advancements in crowd counting algorithms have spurred the development of more refined and specialized datasets with enhanced image quality. Some of the most prominent datasets used to validate these algorithms include JHU-CROWD++ [21], NWPU-Crowd [22], ShanghaiTech [18], UCSD [14], Mall [23], and UCF_CC_50 [24]. Datasets like Mall, UCSD, and ShanghaiTech B are particularly tailored for sparse crowds in single scenes, while JHU-CROWD++, NWPU-Crowd, ShanghaiTech A, and UCF_CC_50 are designed to handle dense crowds in complex environments. Additionally, specialized datasets such as

DLR-ACD, DISCO, and Fudan-ShanghaiTech, which focus on multi-scene aerial images, audiovisual conditions, and video-based counting, respectively, offer targeted research in crowd counting. These resources are indispensable for the validation and development of crowd counting algorithms in diverse scenarios.

UCSD [14]: The UCSD dataset, one of the pioneering datasets for crowd counting, comprises 2000 frames captured from a pedestrian surveillance video. Each frame has a resolution of 238 × 158 pixels, with a total of 49,885 annotated pedestrians. This dataset features a single scene and location, providing a controlled testing environment.

Mall [23]: The Mall dataset, derived from shopping mall surveillance footage, includes 2000 frames, each at a resolution of 320 × 230 pixels. The dataset contains 6000 annotated pedestrians and presents a wider variety of scenes and crowd densities compared to the UCSD dataset.

UCF_CC_50 [24]: This high-density dataset contains 50 images from various scenes such as concert halls and sports arenas, with an average of 1280 people per image, totaling 63,075 annotated heads. The diversity and density of the scenes make it a benchmark for evaluating performance in complex scenarios.

JHU-Crowd++ [21]:This dataset includes 4372 images sourced from the web, with a total of 1.51 million annotations. It features challenging environmental conditions, such as snow and rain, enhancing the diversity of the scenes and presenting additional difficulties for model validation.

NWPU-Crowd [22]: As the largest crowd counting dataset, NWPU-Crowd comprises 5109 images from urban settings, with 2,133,375 annotations. The dataset includes diverse lighting conditions, crowd densities, and negative samples, which contribute to improving the robustness of crowd counting models.

ShanghaiTech [18]: This dataset is divided into two parts. Part A includes 482 images sourced from the web, featuring 241,677 annotations, while Part B contains 716 street images from Shanghai, focusing on lower-density scenes, with 88,488 annotated targets.

To further highlight the contributions of the ClassRoom-Crowd dataset, we provide Table 1 contrasting it with existing datasets like JHU-Crowd++, NWPU-Crowd, and UCSD. Several unique challenges set ClassRoom-Crowd apart from these datasets. Notably, ClassRoom-Crowd features dense occlusions caused by objects such as computer screens and classroom furniture, presenting substantial visual obstructions that make crowd detection and counting significantly more challenging. Moreover, the dataset includes images captured under extreme lighting conditions, such as low-light and uneven illumination, which are less common in other crowd counting datasets. Additionally, the controlled classroom environment, with its varied activities and dynamic crowd configurations, poses distinct challenges compared to the more generalized public scenes found in other datasets. These unique characteristics make ClassRoom-Crowd a more demanding and realistic benchmark, pushing the boundaries of current crowd counting algorithms and fostering advancements in handling occlusions, lighting conditions, and environmental variability.

Table 1. A list of crowd datasets, where "single/variety" refers to whether the data collection scene is singular or diverse. "Normal/challenging" indicates the complexity of the scene. *N* and *M* represent the number of images and the number of annotated objects, respectively.

Dataset	*N*	*M*	Avg. Resolution	Source	Scene	Diversity
UCSD	2000	49,885	158 × 238	Video	Single	Normal
Mall	2000	62,325	480 × 640	Video	Single	Normal
UCF_CC_50	50	63,974	2888 × 2101	Internet	Variety	Normal
JHU-Crowd++	4372	1,515,005	910 × 1430	Internet	Variety	Challenging
NWPU-Crowd	5109	2,133,238	2311 × 3383	Internet	Variety	Challenging
Shanghai Tech PartA	482	241,677	589 × 868	Internet	Variety	Normal
Shanghai Tech PartB	716	88,488	768 × 1024	Video	Single	Normal
ClassRoom-Crowd	7571	172,898	2560 × 1024	Video	Single	Challenging

3. ClassRoom-Crowd Dataset

3.1. Data Collection and Processing

The dataset originates from video recordings captured by surveillance cameras installed in two classrooms, providing both front and rear views. The cameras used are Hikvision's 4-megapixel AcuSense DarkFighter Network Turret Cameras (model: DS-2CD2346TEFWDA4-LS), which offer a resolution of 2560 × 1440 pixels. This dataset encompasses 90 class sessions that document students' movements into and out of the classrooms during class times. While all footage pertains to classroom settings, the videos exhibit variability in timing and lighting conditions. Notably, the dataset includes two segments where students navigate the classroom in complete darkness, highlighting activities under such conditions. The videos were segmented into frames for analysis, and images containing more than five students were annotated to generate corresponding crowd density maps.

3.2. Dataset Annotation

MATLAB R2020a was utilized to annotate individuals in the collected images, with the coordinates of the target individuals recorded to generate ground truth files. These files adhere to the format employed in the UCF and ShanghaiTech datasets. The annotation process was conducted entirely manually, involving five annotators who cross-annotated shuffled images. Images exhibiting significant discrepancies in annotations underwent re-annotation by three annotators. Subsequently, the images were organized sequentially according to the video frames. Observations regarding the number of individuals in the videos were utilized to further enhance annotation accuracy.

3.3. Dataset Characteristics

Our dataset comprises 7571 images with a resolution of 2560 × 1024 pixels and includes 172,898 annotated instances. Figure 1 presents sample images captured from various angles and lighting conditions, along with their corresponding crowd density maps. The dataset is categorized into two parts based on lighting: PART A consists of 7329 images under normal illumination, while PART B contains 242 images taken in low-light or dark environments.

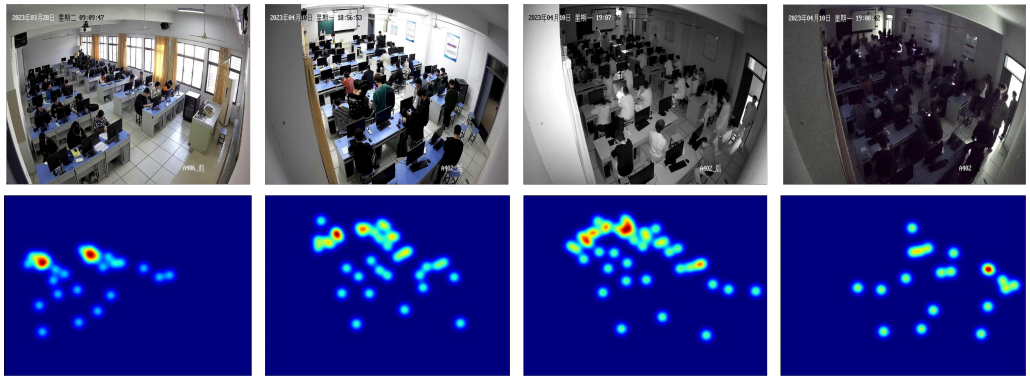


Figure 1. Images and crowd density maps from various perspectives and environments in the Classroom-Crowd dataset.

Our dataset, which include a large collection of images, supports more robust model training by enhancing predictive accuracy and generalization capabilities while minimizing the risk of overfitting. This makes it as an effective benchmark for evaluating and improving the robustness of crowd counting models. Moreover, the dataset provides three distinct advantages:

1. Derived from video frame segmentation, this dataset exhibits temporal continuity, a crucial feature for realistic crowd counting scenarios that are often dynamically

continuous. This characteristic strengthens the relevance of studying temporally sequential videos or images in crowd counting tasks. Compared to datasets derived from web searches, our dataset provides more realistic data for research by offering greater continuity and correlation between frames.

2. This dataset was collected in a computer lab classroom, where computer screens on desks present significant occlusion challenges for crowd counting. The degree of occlusion varies with different individual postures, and the color similarity between heads and the computer screens further complicates crowd counting accuracy.
3. The dataset captures a range of lighting conditions within the computer lab classroom, extending beyond typical indoor illumination. The classroom lighting varies with weather conditions and includes windows facing corridors and outdoors, resulting in uneven light distribution compared to outdoor environments. Notably, the dataset includes images taken in complete darkness, simulating scenarios where students are transitioning between classes or interacting in the classroom. This diversity in lighting provides a crucial test for validating crowd counting methods under extreme conditions.

4. Experiment

In this section, we trained four open-source methods on our proposed dataset and conducted cross-validation tests across various other datasets. Additionally, we evaluated models trained on other datasets using mainstream methods when applied to our dataset. This was followed by a comprehensive analysis and discussion of visualizations derived from the experimental results.

4.1. Baseline Methods

For the experiment, we used three state-of-the-art models [4,25,26] to establish baseline results for our new dataset. These models are summarized as follows:

BL [4]: This method employs a novel Bayesian loss function to address inaccuracies in crowd density maps due to occlusions, perspective changes, and variations in object shapes. The loss function helps generate density maps from point annotations, without imposing constraints on pixel values.

DM-Count [25]: DM-Count introduces a distribution-matching approach for crowd counting that differs from traditional Gaussian smoothing techniques. It leverages optimal transport to assess the similarity between normalized predicted and ground truth density maps, while using total variation loss to stabilize the optimal transport computations.

MAN [26]: MAN is the first to apply Transformer models to crowd counting, using Local Receptive Attention (LRA) and Local Attention Refinement (LAR) modules to manage scale discrepancies that CNNs and global attention mechanisms struggle to address. It also incorporates specialized loss functions to mitigate label noise.

4.2. Experiment Protocol and Analysis

We defined three experimental protocols to generate baseline results: (1) Protocol A (cross-environment): PART A was used for training, and PART B was employed for testing, following mainstream practices. (2) Protocol B (mixed evaluation): Both PART A and PART B were used for training, with PART B reserved for testing. PART B was divided into a training set (80%) and a testing set (20%). The training portions of PART A and PART B were combined to form the training set for Protocol B, while the testing set from PART B served as the evaluation set. (3) Protocol C (cross-dataset): Existing mainstream datasets were used for training, while PART A, PART B, and the combination of PART A and PART

B were used as test sets. The experimental results of Protocol A, Protocol B, and Protocol C are shown in Table 2 and Table 3, respectively.

Table 2. Experimental results of Protocol A and Protocol B.

Method	Protocol A		Protocol B	
	MAE	MSE	MAE	MSE
BL [4]	1.49	1.94	10.42	11.1
DM-Count [25]	1.59	2.13	10.94	11.74
MAN [26]	35.26	36.32	36.23	36.84

Table 3. Experimental results of Protocol C.

Test Set		PART A		PART B		PART A + PART B	
Method	Train Set	MAE	MSE	MAE	MSE	MAE	MSE
BL (ICCV 19)	SHHA	12.99	13.69	17.06	28.27	6.22	9.48
	SHHB	8.62	13.86	16.70	17.31	8.93	14.01
	UCF	5.96	9.28	12.99	13.69	6.22	9.48
DM-Count (NIPS20)	SHHA	4.28	5.61	21.24	21.86	4.93	6.96
	SHHB	5.25	6.96	25.76	26.59	6.03	8.57
MAN (CVPR22)	JHU	6.78	8.11	19.58	20.16	7.26	8.87
	UCF	3.86	5.60	9.53	10.58	4.07	5.87

A comparison of results between Protocols A and B, as shown in Figure 2, indicates that error values for all methods under Protocol B are consistently lower than those under Protocol A. This demonstrates a notable enhancement in the models' cross-domain detection capabilities, achieved by incorporating images from extreme conditions into the training set. To further improve the generalization ability of crowd counting models in complex environments, it is crucial to diversify the training conditions.

Additionally, Figure 2 illustrates that models designed for large-scale crowd counting, such as MAN, exhibit significantly higher detection errors in sparse scenes compared to traditional methods. This underscores a limitation in current algorithms, which struggle to effectively manage crowd counting tasks across both sparse and dense environments.

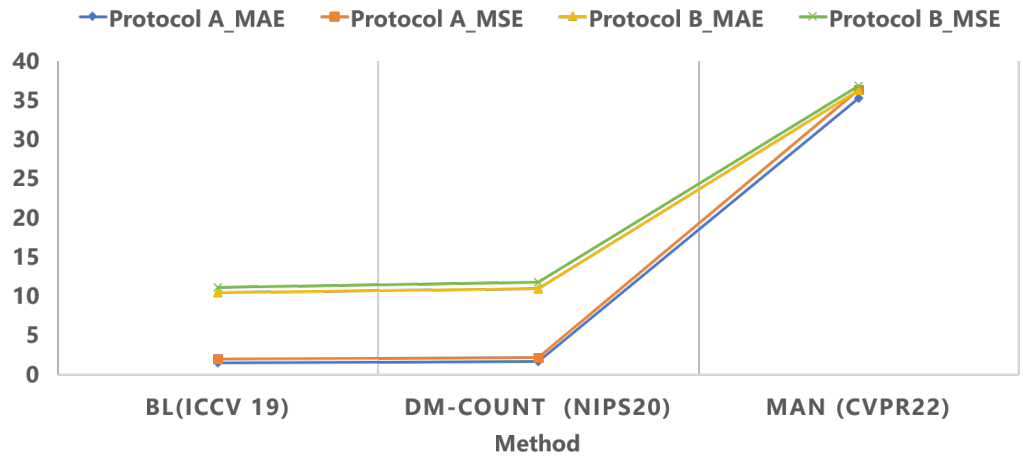


Figure 2. Protocol A and Protocol B use different methods to train the model and get results as shown in the figure. MAE is the mean absolute error and MSE is the mean square error.

Moreover, results from Protocol C reveal that models trained on various datasets, regardless of whether they employ conventional or advanced methods, experience a marked decline in performance under extreme conditions compared to normal conditions. This finding highlights the need for continued research into crowd counting under challenging scenarios, such as low-light environments, to address cross-domain issues and improve model robustness.

5. Conclusions and Future Perspectives

This paper introduces a crowd counting dataset specifically designed for classroom environments, characterized by variations in crowd density under extreme conditions, temporal continuity, and high resolution. The dataset provides essential support for addressing crowd counting problems under challenging conditions and for conducting temporal analyses in crowd counting tasks.

Our experimental analysis has highlighted several areas for future improvement in enhancing crowd counting models. Ensuring model stability in extreme weather and environmental conditions is crucial, and incorporating negative samples is necessary to enhance robustness. Furthermore, since real-world crowd counting encompasses both sparse and dense scenes, additional research is required to ensure model accuracy across these variations. Addressing occlusion challenges is also essential, as real-world counting often involves obstructed spaces. Moreover, utilizing temporal data from continuous scenes can improve real-time counting accuracy, necessitating faster data processing and real-time analysis capabilities in future models. These enhancements are critical for developing more effective crowd counting techniques.

While the ClassRoom-Crowd dataset primarily focuses on static frames, it presents significant opportunities for deeper temporal analysis in future research. As the dataset is derived from sequential video footage, consecutive frames inherently maintain temporal continuity. In future work, we aim to expand the dataset's utility by incorporating methods that explicitly leverage this temporal information. For instance, integrating temporal models such as Long Short-Term Memory (LSTM) networks, 3D Convolutional Neural Networks (3D-CNNs), or optical flow techniques could enhance a model's ability to track changes in crowd density, detect movement patterns, and better manage occlusions over time. This approach could substantially improve the detection accuracy and robustness of crowd counting algorithms in dynamic environments, making the dataset even more valuable for real-time applications such as surveillance and behavior analysis.

Author Contributions: Conceptualization, X.H.; methodology, W.J.; software, W.J.; validation, Q.Z and S.L.; formal analysis, Q.Z.; investigation, X.H.; resources, X.H.; data curation, W.J.; writing—original draft preparation, W.J.; writing—review and editing, X.H.; visualization, Q.Z.; supervision, X.H.; project administration, X.H.; funding acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Natural Science Foundation of China under Grant 62076122, Research funding of NJIT (No. YKJ201982), the Fundamental Research Funds for the Central Universities (No. 2242024k30027), and Basic Science (Natural Science) research project of higher education institutions in Jiangsu Province (24KJA520003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent for participation is not required as per local legislation (Nanjing Institute of Technology).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SIFT	Scale-Invariant Feature Transform
LBP	Local Binary Pattern
DM-Count	Distribution matching for crowd counting
BL	Bayesian loss for crowd count estimation with point supervision
MAN	Boosting crowd counting via multifaceted attention.

References

1. Cardia, M.; Luca, M.; Pappalardo, L. Enhancing crowd flow prediction in various spatial and temporal granularities. In Proceedings of the Companion Proceedings of the Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 1251–1259.
2. Sun, Z.; Chen, J.; Chao, L.; Ruan, W.; Mukherjee, M. A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 1819–1833. [CrossRef]
3. Ma, Z.; Hong, X.; Wei, X.; Qiu, Y.; Gong, Y. Towards a universal model for cross-dataset crowd counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 3205–3214.
4. Ma, Z.; Wei, X.; Hong, X.; Gong, Y. Bayesian loss for crowd count estimation with point supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Repubic of Korea, 27 October–2 November 2019; pp. 6142–6151.
5. Liu, W.; Durasov, N.; Fua, P. Leveraging self-supervision for cross-domain crowd counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 5341–5352.
6. Liu, J.; Gao, C.; Meng, D.; Hauptmann, A.G. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5197–5206.
7. Liu, Y.; Shi, M.; Zhao, Q.; Wang, X. Point in, box out: Beyond counting persons in crowds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 6469–6478.
8. Zhou, S.; Wang, J.; Meng, D.; Liang, Y.; Gong, Y.; Zheng, N. Discriminative feature learning with foreground attention for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 4671–4684. [CrossRef] [PubMed]
9. Khan, M.H.; Shirahama, K.; Farid, M.S.; Grzegorzek, M. Multiple human detection in depth images. In Proceedings of the 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP), Montreal, QC, Canada, 21–23 September 2016; pp. 1–6.
10. Liu, B.; Vasconcelos, N. Bayesian model adaptation for crowd counts. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4175–4183.
11. Shang, C.; Ai, H.; Bai, B. End-to-end crowd counting via joint learning local and global count. In Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, Phoenix, AZ, USA, 25–28 September 2016; pp. 1215–1219.
12. Chattopadhyay, P.; Vedantam, R.; Selvaraju, R.R.; Batra, D.; Parikh, D. Counting everyday objects in everyday scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1135–1144.
13. Chen, K.; Gong, S.; Xiang, T.; Change Loy, C. Cumulative attribute space for age and crowd density estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2467–2474.
14. Chan, A.B.; Liang, Z.S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Anchorage, Alaska, 23–28 June 2008; pp. 1–7.
15. Wang, C.; Zhang, H.; Yang, L.; Liu, S.; Cao, X. Deep people counting in extremely dense crowds. In Proceedings of the ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1299–1302.
16. Lempitsky, V.; Zisserman, A. Learning to count objects in images. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 1–9.
17. Fiaschi, L.; Köthe, U.; Nair, R.; Hamprecht, F.A. Learning to count with regression forest and structured labels. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 2685–2688.
18. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 589–597.
19. Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid cnns. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1861–1870.
20. Oghaz, M.M.; Khadka, A.R.; Argyriou, V.; Remagnino, P. Content-aware density map for crowd counting and density estimation. *arXiv* **2019**, arXiv:1906.07258.

21. Sindagi, V.A.; Yasarla, R.; Patel, V.M. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2594–2609. [CrossRef] [PubMed]
22. Wang, Q.; Gao, J.; Lin, W.; Li, X. NWPU-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2141–2149. [CrossRef] [PubMed]
23. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature mining for localised crowd counting. In Proceedings of the BMVC, Surrey, UK, 3–7 September 2012; Volume 1, p. 3.
24. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 532–546.
25. Wang, B.; Liu, H.; Samaras, D.; Nguyen, M.H. Distribution matching for crowd counting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1595–1607.
26. Lin, H.; Ma, Z.; Ji, R.; Wang, Y.; Hong, X. Boosting crowd counting via multifaceted attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 11–20 June 2022; pp. 19628–19637.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.