Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

Crowd Density Estimation via a VGG-16-Based CSRNet Model

Damla Tatlıcan[1], Nafiye Nur Apaydin[1], Orhan Yaman[1*], Mehmet Karakose[2]

[1] Department of Digital Forensics Engineering, College of Technology, Firat University, 23119 Elazig, Turkey
[2] Department of Computer Engineering, Faculty of Engineering, Firat University, 23119 Elazig, Turkey

* Correspondence: Orhan Yaman (orhanyaman@firat.edu.tr)

Received: 03-13-2025 **Revised:** 04-14-2025 **Accepted:** 04-25-2025

Citation: D. Tatlıcan, N. N. Apaydin, O. Yaman, and M. Karakose, "Crowd density estimation via a VGG-16-based CSRNet model," *Inf. Dyn. Appl.*, vol. 4, no. 2, pp. 66–75, 2025. https://doi.org/10.56578/ida040201.

Abstract: Accurate crowd density estimation has become critical in applications ranging from intelligent urban planning and public safety monitoring to marketing analytics and emergency response. In recent developments, various methods have been used to enhance the precision of crowd analysis systems. In this study, a Convolutional Neural Network (CNN)-based approach was presented for crowd density detection, wherein the Congested Scene Recognition Network (CSRNet) architecture was employed with a Visual Geometry Group (VGG)-16 backbone. This method was applied to two benchmark datasets—Mall and Crowd-UIT—to assess its effectiveness in real-world crowd scenarios. Density maps were generated to visualize spatial distributions, and performance was quantitatively evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics. For the Mall dataset, the model achieved an MSE of 0.08 and an MAE of 0.10, while for the Crowd-UIT dataset, an MSE of 0.05 and an MAE of 0.15 were obtained. These results suggest that the proposed VGG-16-based CSRNet model yields high accuracy in crowd estimation tasks across varied environments and crowd densities. Additionally, the model demonstrates robustness in generalizing across different dataset characteristics, indicating its potential applicability in both surveillance systems and public space management. The outcomes of this investigation offer a promising direction for future research in data-driven crowd analysis, particularly in enhancing predictive reliability and real-time deployment capabilities of deep learning models for population monitoring tasks.

Keywords: Crowd density estimation; CSRNet; VGG-16; Mall dataset; Crowd-UIT dataset

1 Introduction

Crowds are predominantly observed in public and private spaces where people interact, engage in leisure activities, travel, and exchange information, such as shopping malls, cinemas, concerts, sports halls, public transportation, traffic, schools, and hospitals. Crowd density varies depending on the number of people present. For instance, in rural areas, due to the limited space and lower population density, crowd intensity remains low. In contrast, in metropolitan cities, the combination of large urban areas and high population density results in significantly higher crowd concentrations. Crowd density detection offers multiple benefits. For example, in areas with high crowd density, early intervention can help prevent security threats, such as public disturbances or fights, by enabling timely precautions. In the event of natural disasters, it facilitates effective search and rescue operations. In traffic management, red light durations can be adjusted based on the presence of vehicles, optimizing traffic flow. In shopping malls, crowd density detection allows for the identification of frequently visited stores, enabling strategic placement of less-visited stores next to high-traffic ones to attract more customers and increase revenue. Additionally, analyzing customer behavior in highly visited stores helps determine which products are frequently purchased together, allowing for strategic shelf placement to enhance marketing strategies. Moreover, crowd density detection can be utilized to monitor gatherings and implement preventive measures to mitigate the spread of infectious diseases. Due to these advantages, crowd density detection remains a crucial research area, with various studies continuing to be conducted in the literature.

Despite the wide range of applications, crowd density detection still faces several challenges, particularly in scenarios involving high crowd density, severe congestion, and varying illumination or perspective conditions. To address these problems, various approaches have been proposed in the literature, ranging from traditional image processing techniques to modern deep learning-based methods.

Table 1 provides a comparative overview of several studies concerning crowd density detection. The methods employed in these studies range from traditional machine learning algorithms, such as random forests and regression models, to deep learning architectures, including CNN-based and attention-based models. While classical models offer advantages in terms of simplicity and interpretability, they often fall short in effectively extracting features from densely populated images.

Table 1. Some studies for crowd density detection

References	Year	Purpose of the Studies	Advantages	Disadvantages
Wei et al. [1]	2019	In this study, a novel algorithm was proposed to address the counting of fastmoving crowds by utilizing deep cumulative feature learning, support vector regression, and spatiotemporal features.	Handles motion features and uses temporal information.	Computationa lly intensive and less effective in static scenes.
Zhang et al. [2]	2015	In this study, a label distribution learning method was introduced for crowd counting in public video surveillance.	Performs well under varying density levels.	Limited performance with perspective distortion.
Chen et al. [3]	2012	In this study, a multi-output regression model was developed for crowd counting in spatially localized regions.	Effective in structured and partitioned scenes.	Less accurate in highly dense scenarios.
Chen et al. [4]	2013	In this study, age and crowd density estimation were performed by transforming high-dimensional feature vectors into low-dimensional scalar values. To achieve this, a cumulative feature space was utilized to reduce data imbalance.	Handles highdimensional data and reduces imbalance.	Information loss during reduction and sensitivity to feature selection.
Pham et al. [5]	2015	In this study, a random forest-based model was developed to estimate crowd density in scenes with dense human crowds.	Robust to noise and outliers and interpretable.	May underperform in real-time applications; limited spatial modeling.
Tomar et al. [6]	2022	In this study, a dynamic kernel-based CNN-Linear Regression (LR) model was proposed for human counting in crowd scenes. This model is specifically designed to address the overlapping issues in dense crowds.	Tackles overlap issues; dynamic kernel improves learning.	Requires large datasets for training and complex parameter tuning.
Abdullah & Jalal [7]	2023	In this study, a fuzzy classifier (neurofuzzy classifier) and a semantic segmentation-based method were developed for crowd tracking and anomaly detection in intelligent surveillance systems. The method aims to monitor behaviors in crowded areas and detect anomalies.	Interpretable decision-making; good for anomaly detection.	Sensitive to noise; limited scalability in large scenes.
Tripathy & Srivastava [8]	2021	In this study, an Attention-based MultiStream (AMSCNN)-CNN was developed for video-based crowd counting, considering both spatial and temporal features.	Captures both spatial and temporal features effectively.	High computational cost; real-time deployment is challenging.

References	Year	Purpose of the Studies	Advantages	Disadvantages
Xiong et al. [9]	2017	In this study, a model was developed for crowd counting in videos that captures temporal and spatial information, enabling more accurate predictions.	Improves accuracy in dynamic scenes.	Limited performance in still images; temporal data dependency.
Maktoof et al. [10]	2023	In this study, different models from the You Only Look Once version 5 (YOLOv5) family were compared to detect human crowds more accurately.	High detection accuracy and real-time capability.	Struggles in extremely dense scenes; limited counting ability.
Maktoof et al. [11]	2023	In this study, a real-time system combining YOLOv5 and Kernel Correlation Filter (KCF) algorithms was developed for detecting and counting human crowds.	Efficient realtime tracking and detection.	Accuracy drops in highly occluded scenes.
Deng et al. [12]	2024	In this study, a new dataset was presented in the field by performing human crowd counting from videos.	Enriches the field with new data sources.	No new method proposed; limited to data contribution.

Conventional CNN-based models often fail to effectively extract features from highly congested scenes, resulting in poor performance in detecting individuals within dense crowds. Therefore, in this study, CSRNet, a CNN architecture specifically designed for crowd counting and density map estimation, was proposed. The proposed method integrates a VGG-16-based frontend with an extended convolutional backend, enabling CSRNet to capture multi-scale spatial features effectively without losing resolution. This allows for accurate localization and counting of individuals even in complex and densely crowded scenarios. The primary objective of this study is to develop a fast and accurate crowd density detection model capable of robustly estimating the number and positions of people in an image, particularly under challenging conditions. Another aim of this work is to contribute to the advancement of the existing CSRNet model in the literature. Moreover, due to the limited number of existing studies utilizing the Crowd-UIT dataset, this work is expected to serve as a valuable reference for future research efforts.

2 Methodology

2.1 Dataset

In this study, the Mall and Crowd-UIT datasets from the literature were used [13].

2.1.1 Mall dataset

The Mall dataset consists of low-resolution, crowded human images captured in a shopping mall. This dataset contains 2,000 images, with the number of people in each image provided in a CSV file along with the image name and corresponding person count. Figure 1 presents sample images from the Mall dataset along with the CSV file.

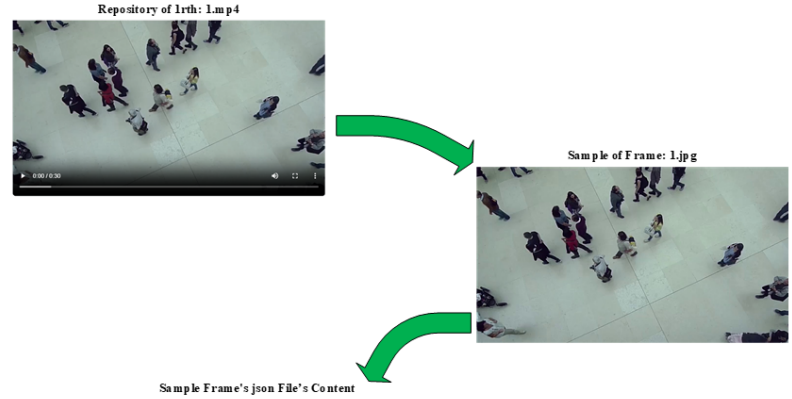
	ID	Count
0	Seq_000001.jpg	35
1	Seq_000002.jpg	41
2	Seq_000003.jpg	41
3	Seq_000004.jpg	44

(b)

Figure 1. Mall dataset (a) Original images (b) CSV file

2.1.2 Crowd-UIT dataset

The Crowd-UIT dataset consists of ten videos [12]. Using these videos, ten different repositories have been created for each video. In this study, only repositories 1 and 2 from the Crowd-UIT dataset were used. Each repository contains frames from the videos and the locations of people in those frames as JSON labels. The first repository contains 145 images, while the second repository contains 144 images. Figure 2 presents the sample image and the JSON file for repository 1, while Figure 3 presents the sample image and the JSON file for repository 2.



Sample Frame's json File's Content

"root":[24 items 0:[3 items 0:int1067 1:int332 2:int1]	1:[3 items 0:int486 1:int384 2:int2]	2:[3 items 0:int588 1:int282 2:int3]	3:[3 items 0:int207 1:int277 2:int4]	4:[3 items 0:int212 1:int322 2:int5]	5:[3 items 0:int365 1:int267 2:int6]
6:[3 items 0:int408 1:int208 2:int7]	7:[3 items 0:int395 1:int113 2:int8]	8:[3 items 0:int318 1:int182 2:int9]	9:[3 items 0:int13 1:int648 2:int10]	10:[3 items 0:int11 1:int115 2:int11]	11:[3 items 0:int9 1:int24 2:int12]
12:[3 items 0:int101 1:int5 2:int13]	13:[3 items 0:int150 1:int10 2:int14]	14:[3 items 0:int568 1:int83 2:int15]	15:[3 items 0:int664 1:int17 2:int16]	16:[3 items 0:int698 1:int6 2:int17]	17:[3 items 0:int733 1:int9 2:int18]
18:[3 items 0:int827 1:int9 2:int19]	19:[3 items 0:int1263 1:int470 2:int20]	20:[3 items 0:int1258 1:int165 2:int21]	21:[3 items 0:int1262 1:int594 2:int22]	22:[3 items 0:int1253 1:int695 2:int23]	23:[3 items 0:int744 1:int207 2:int24]]

Figure 2. Sample image for repository 1 of the Crowd-UIT dataset and corresponding JSON tag

2.2 Method

The proposed method in this study was applied separately to the repositories in the Mall and Crowd-UIT datasets to obtain results. After training, the resulting model was tested on sample images. The block diagram of the proposed method is shown in Figure 4.

As shown in Figure 4, the images and labels from the dataset were first fed as input to the VGG-16 model. VGG-16 was used to extract the basic features and subsequently generate feature maps. These feature maps were then passed to the convolutional layers of CSRNet, where more complex operations were performed to detect the density [14–20]. Following this, the information from a larger area was captured by passing the feature maps through dilated convolution layers. After these steps, density maps were generated. The resulting density maps were then upsampled to higher resolution through upsampling operations. To determine the accuracy of the model and compare it with other studies, MSE and MAE metrics were calculated. These values were used to assess the model's learning process and performance.

The MSE and MAE metrics are given in Eq. (1) and Eq. (2):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_i - X| \tag{1}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_i)^2 \tag{2}$$

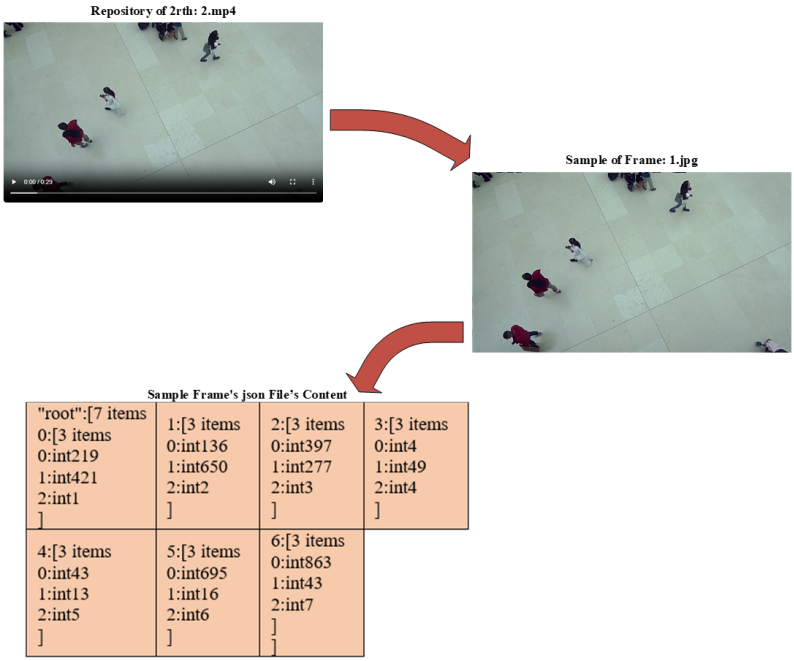


Figure 3. Sample image for repository 2 of the Crowd-UIT dataset and corresponding JSON tag

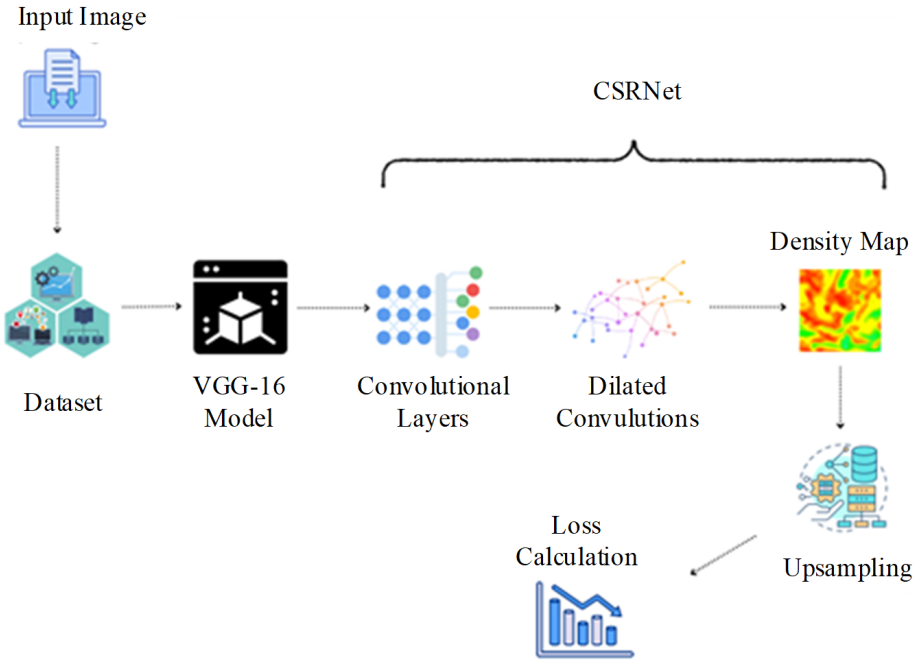


Figure 4. Block diagram of the proposed CSRNet model

2.2.1 MDensity maps

Density maps visualize the concentrations of people in images, clearly showing which areas of the crowd are more densely populated. The model predicts the human density in specific areas of the given image and provides this as a density map. These maps are generated through convolutional operations via deep learning models. In the maps, blue colors represent areas with low density. As the density increases, the colors transition to yellow and orange. The regions with the highest crowd density are shown in red. The working principle of the density map is illustrated in Figure 5.

CSRNet is a deep learning model specifically designed for estimating crowd density. Its most distinctive feature lies in its utilization of dilated convolutional layers, which enable effective crowd counting, particularly in densely populated scenarios. In this study, the VGG-16 architecture was employed as the frontend of the CSRNet model.

VGG-16 is a pre-trained deep CNN, originally trained on the ImageNet dataset, and is widely used for feature extraction due to its strong generalization capability. For the purposes of this study, only the convolutional layers of VGG-16 were retained, while the fully connected layers were removed. This modification allows the model to accept input images of arbitrary sizes and produce corresponding density maps, making it suitable for practical crowd analysis applications. The convolutional blocks in VGG-16 extract hierarchical features, where the early layers learn low-level visual information such as edges and textures, while the deeper layers capture high-level semantic features. This progressive feature representation enhances the model's capability in recognizing and interpreting complex crowd patterns. The architecture of the modified VGG-16 frontend and CSRNet backend is illustrated in Figure 6.

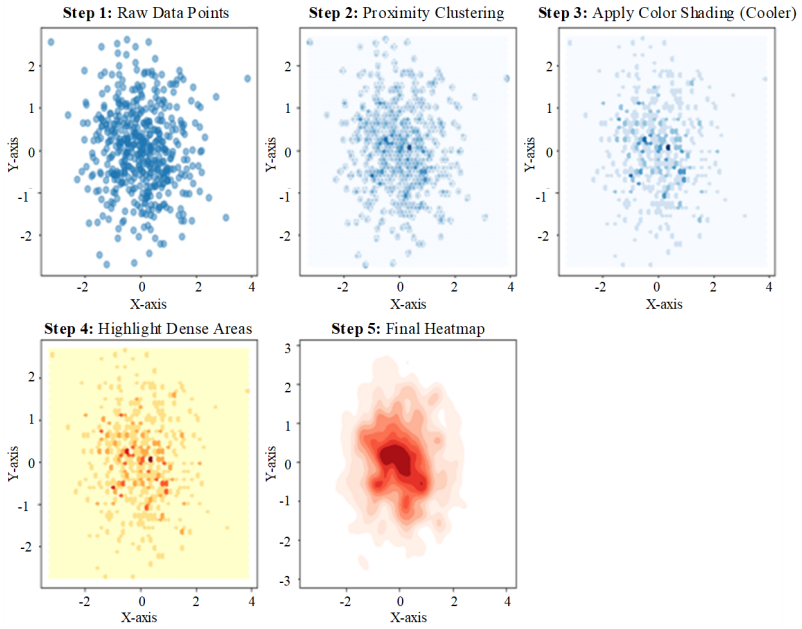


Figure 5. Working principle of the density map

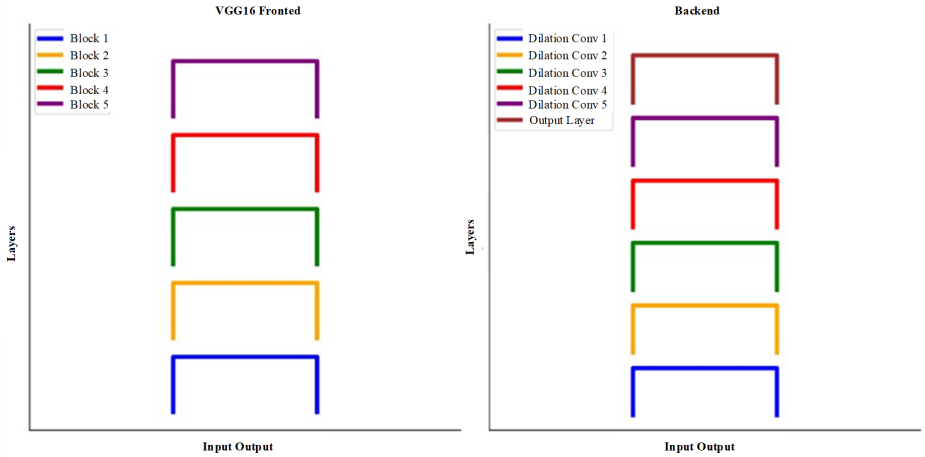


Figure 6. Architecture of the VGG-16 model

The CSRNet backend comprises five consecutive dilated convolutional layers, each designed to expand the receptive field without reducing the feature map resolution. By increasing the dilation rate, these layers effectively aggregate more contextual information while maintaining the integrity of fine-grained spatial details—a critical factor in the dense crowd analysis.

In addition to these core components, the following layers were utilized:

- Convolutional layers: Feature maps were extracted from the input by applying a set of learnable filters. The size of these filters determines the level of detail captured in the feature maps.
- MaxPooling layers: The feature maps were downsampled by selecting the maximum value from subregions, thus preserving the most salient features while reducing computational complexity.

- Sequential layer configuration: Layers were structured sequentially to streamline the architectural design, allowing for organized and efficient forward propagation.

This enhanced architecture allows the model to process images of varying sizes and to make accurate and real-time predictions even in complex and crowded environments. All these layers and the input-output values are shown in Figure 7.

Figure 7 provides a comprehensive visualization of the layer-by-layer architecture of the CSRNet model enhanced with VGG-16 as the frontend. The diagram outlines the exact input and output dimensions of each layer, starting from the input layer and continuing through the convolutional and pooling blocks of VGG-16, followed by the dilated convolutional layers and the final output layer. The frontend, consisting of five convolutional blocks (Block1 to Block5), is derived from VGG-16 and is responsible for feature extraction. These blocks utilize multiple Conv2D layers followed by MaxPooling2D layers, which progressively reduce the spatial resolution while increasing the depth (i.e., number of feature maps), capturing both low- and high-level features effectively. After the VGG-16 blocks, the backend (CSRNet) begins with a set of dilated convolutional layers (represented in the model with three Conv2D layers post-Block5), which preserve spatial resolution while expanding the receptive field. This allows the model to integrate broader contextual information for precise density estimation. The final sequential layer produces a single-channel output representing the predicted density map. This detailed structure illustrates how the model maintains a balance between spatial accuracy and semantic richness, enabling robust performance in dense crowd counting tasks.

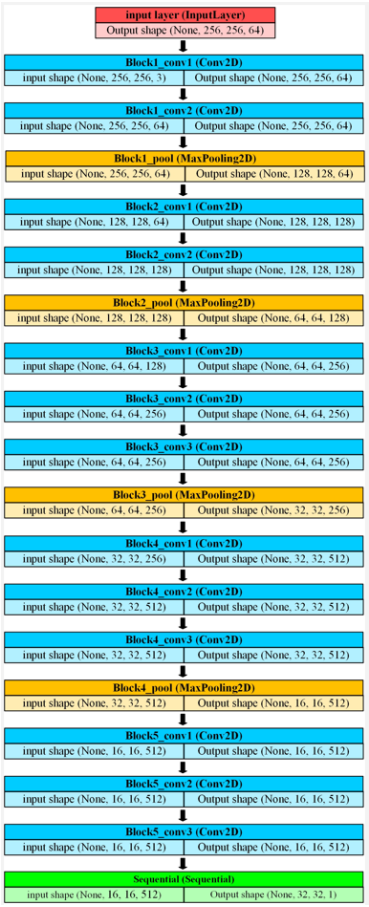


Figure 7. Layers and input-output values of the CSRNet model enhanced with VGG-16

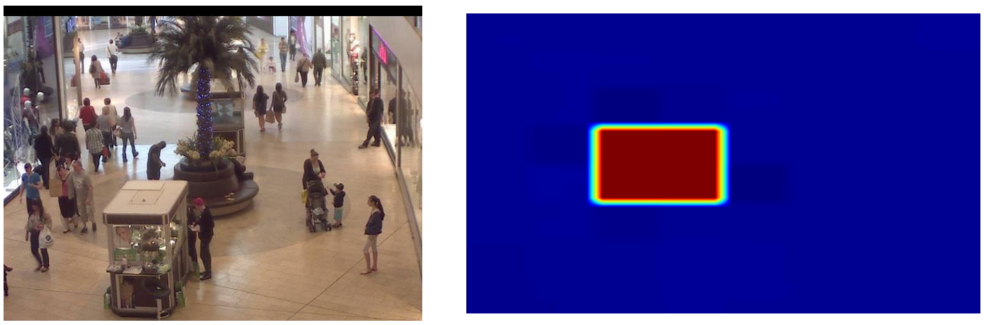
3 Results

3.1 Mall dataset results

Figure 8 presents a test example of the model created using the Mall dataset.

This dataset only provides the number of people along with the images and does not give the locations of the people. Therefore, the model attempts to predict the locations of people on its own.

The results of the proposed method for the MALL dataset were compared with the studies in the literature in Table 2.



(a) (b)

Figure 8. Test example on the MALL dataset (a) Original image (b) Density detection image

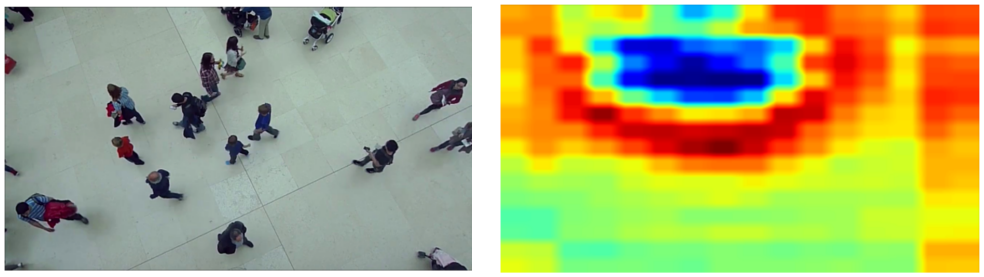
Table 2. Comparison of the MALL dataset results with existing studies

References	Methods	MSE	MAE
Wei et al. [1]	Boosting DAL-SVR	9.57	2.40
Zhang et al. [2]	IIS-LDL	12.1	2.69
Chen et al. [3]	Multi-output ridge regression (MORR)	15.7	3.15
Chen et al. [4]	CA-RR	17.7	3.43
Pham et al. [5]	Patch-based Gaussian kernel	10.0	2.50
Tomar et al. [6]	DKCNN-LR	2.76	1.65
Abdullah & Jalal [7]	Semantic segmentation-based crowd tracking method	4.34	2.57
Tripathy & Srivastava [8]	AMS-CNN	3.08	2.47
Xiong et al. [9]	ConvLSTM Bidirectional ConvLSTM	7.6	2.10
VGG-16 + CSRNet (proposed model)		**0.08**	**0.10**

Upon examining Table 2, it can be seen that the CSRNet model enhanced with VGG-16 achieves significant success in crowd density prediction, with an MSE of 0.08 and an MAE of 0.10, outperforming other methods in the literature.

3.2 Crowd-UIT dataset results

Figure 9 shows the test result with an example image after applying the proposed method to the Crowd-UIT dataset.



(a) (b)

Figure 9. Test example on the Crowd-UIT dataset (a) Original image (b) Density detection image

The Crowd-UIT dataset not only contains images of crowded scenes but also includes precise location annotations for each individual. These annotations were directly utilized during the training phase of the proposed method, enabling more precise learning of spatial patterns associated with crowd distribution. By incorporating positional labels into the supervised learning process, the model could effectively minimize the loss function by explicitly correlating input image features with ground-truth spatial positions. This significantly improved the optimization process during training, as the model received clear, structured signals about where people are located, rather than only how many people are in the scene. As a result, it could better localize individuals even in densely populated scenes or in areas with high background complexity.

Table 3 presents a comparative analysis of the Crowd-UIT dataset results with existing studies in the literature. As shown, there are very few prior works utilizing the Crowd-UIT dataset, and all the known studies are summarized in Table 3. Due to this scarcity, the findings of this study are anticipated to contribute as a meaningful benchmark for future researchers working on crowd analysis.

Table 3. Comparison of the Crowd-UIT dataset results with existing studies

References	Methods	MSE	MAE	Accuracy
Deng et al. [12]	FairMOT	5232.8	50.4	-
	JDE	439425.5	445.5	
Maktoof et al. [10]	YOLOv5s	-	-	93.22%
	YOLOv5m			90.96%
	YOLOv51			96.41%
	YOLOv5x			96.53%
Maktoof et al. [11]	YOLOv5 + KCF	-	-	97.61%
VGG-16 + CSRNet (proposed model)		**0.05**	**0.15**	-

Upon examining Table 3, it can be seen that the CSRNet model enhanced with VGG-16 achieves significant success in crowd density prediction, with an MSE of 0.05 and an MAE of 0.15, outperforming other studies in the literature.

In conclusion, the Crowd-UIT dataset includes crowd images along with annotated human positions. Since the proposed method was trained with these positional annotations, it was able to more accurately detect individuals regardless of crowd density. In contrast, the Mall dataset only provides the total number of people in each image without their exact positions. Therefore, the proposed method had to autonomously identify individuals in the scenes. However, as illustrated in the test image shown in Figure 8, due to the high scene complexity in the Mall dataset—such as lighting angles and the presence of non-human objects like trees—the method was not as successful as it was on the Crowd-UIT dataset. Therefore, when compared to the Mall dataset, the Crowd-UIT dataset yielded better results due to its higher image clarity and the availability of directly annotated human positions.

4 Conclusions

In this study, crowd density detection was performed using the CSRNet model enhanced with VGG-16 on the Mall and Crowd-UIT datasets. The model was applied separately to each dataset, and the results were obtained. The density maps generated by the model effectively visualized areas with high crowd density. The MSE and MAE metrics obtained in the model showed very low values. For the Mall dataset, an MSE of 0.08 and an MAE of 0.10 were achieved. For the Crowd-UIT dataset, an MSE of 0.05 and an MAE of 0.15 were obtained. When comparing the Crowd-UIT dataset with the Mall dataset, better results were observed in the Crowd-UIT dataset due to the clearer images and the direct labeling of people's positions. The low error values provided by the CSRNet model enhanced with VGG-16 demonstrate that the model performs better than existing methods in the literature for crowd density detection. Since other algorithms have higher error metrics, it is evident that the CSRNet model enhanced with VGG-16 is an effective method for crowd density detection.

Author Contributions

The authors contributed equally to the article.

Funding

The paper was funded by the Scientific and Technological Research Council of Turkey (Grant No.: 5220154).

Data Availability

1. The data [image] supporting our research results are deposited in [Mall Dataset], which does not issue DOIs. The data can be accessed at [https://personal.ie.cuhk.edu.hk/ ccloy/downloads_mall_dataset.html].

2. The data [image] supporting our research results are deposited in [Crowd-UIT], which does not issue DOIs. The data can be accessed at [https://www.kaggle.com/datasets/khitthanhnguynphan/crowduit?select=Crowd-UIT].

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] X. L. Wei, J. P. Du, M. Y. Liang, and L. F. Ye, "Boosting deep attribute learning via support vector regression for fast moving crowd counting," *Pattern Recognit. Lett.*, vol. 119, pp. 12–23, 2019. https://doi.org/10.1016/j.patrec.2017.12.002

[2] Z. X. Zhang, M. Wang, and X. Geng, "Crowd counting in public video surveillance by label distribution learning," *Neurocomputing*, vol. 166, pp. 151–163, 2015. https://doi.org/10.1016/j.neucom.2015.03.083

[3] K. Chen, C. C. Loy, S. G. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *BMVC*, 2012, pp. 1–11.

[4] K. Chen, S. G. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, *Portland, OR, USA*, 2013, pp. 2467–2474. https://doi.org/10.1109/CVPR.2013.319

[5] V. Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, *Santiago, Chile*, 2015, pp. 3253–3261. https://doi.org/10.1109/ICCV.2015.372

[6] A. Tomar, S. Kumar, B. Pant, and U. K. Tiwari, "Dynamic kernel CNN-LR model for people counting," *Appl. Intell.*, vol. 52, no. 1, pp. 55–70, 2022. https://doi.org/10.1007/s10489-021-02375-6

[7] F. Abdullah and A. Jalal, "Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system," *Arab. J. Sci. Eng.*, vol. 48, no. 2, pp. 2173–2190, 2023. https://doi.org/10.1007/s13369-022-07092-x

[8] S. K. Tripathy and R. Srivastava, "AMS-CNN: Attentive multi-stream CNN for video-based crowd counting," *Int. J. Multimed. Inf. Retr.*, vol. 10, no. 4, pp. 239–254, 2021. https://doi.org/10.1007/s13735-021-00220-7

[9] F. Xiong, X. J. Shi, and D. Y. Yeung, "Spatiotemporal modeling for crowd counting in videos," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5151–5159.

[10] M. A. J. Maktoof, I. T. A. Al_attar, and I. N. Ibraheem, "Comparison YOLOv5 family for human crowd detection," *Int. J. Online Biomed. Eng.*, vol. 19, no. 4, pp. 94–108, 2023. https://doi.org/10.3991/ijoe.v19i04.39095

[11] M. A. J. Maktoof, I. N. Ibraheem, and I. T. A. Al_attar, "Crowd counting using Yolov5 and KCF," *Period. Eng. Nat. Sci.*, vol. 11, no. 2, pp. 92–101, 2023. https://doi.org/10.21533/pen.v11.i2.102

[12] L. J. Deng, Q. H. Zhou, S. H. Wang, J. M. Górriz, and Y. D. Zhang, "Deep learning in crowd counting: A survey," *CAAI Trans. Intell. Technol.*, vol. 9, no. 5, pp. 1043–1077, 2024. https://doi.org/10.1049/cit2.12241

[13] C. Change Loy, S. G. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *Proceedings of the IEEE International Conference on Computer Vision*, *Sydney, NSW, Australia*, 2013, pp. 2256–2263. https://doi.org/10.1109/ICCV.2013.270

[14] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," in *2019 International Conference on Information Technology (ICIT)*, *Bhubaneswar, India*, 2019, pp. 94–98. https://doi.org/10.1109/ICIT48102.2019.00023

[15] Y. H. Li, X. F. Zhang, and D. M. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Lake City, UT, USA*, 2018, pp. 1091–1100. https://doi.org/10.1109/CVPR.2018.00120

[16] C. L. Li, T. J. N. Yang, S. J. Zhu, C. Chen, and S. Y. Guan, "Density map guided object detection in aerial images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, *Seattle, WA, USA*, 2020, pp. 737–746. https://doi.org/10.1109/CVPRW50498.2020.00103

[17] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for classifying images," *Int. J. Sci. Res. Publ.*, vol. 9, no. 10, pp. 143–150, 2019. https://doi.org/10.29322/IJSRP.9.10.2019.p9420

[18] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, *Seoul, Korea (South)*, 2019, pp. 1130–1139. https://doi.org/10.1109/ICCV.2019.00122

[19] J. J. Xiong, L. M. Po, W. Y. Yu, C. Zhou, P. F. Xian, and W. F. Ou, "CSRNet: Cascaded selective resolution network for real-time semantic segmentation," *Expert Syst. Appl.*, vol. 211, p. 118537, 2023. https://doi.org/10.1016/j.eswa.2022.118537

[20] C. Sitaula and M. B. Hossain, "Attention-based VGG-16 model for COVID-19 chest X-ray image classification," *Appl. Intell.*, vol. 51, no. 5, pp. 2850–2863, 2021. https://doi.org/10.1007/s10489-020-02055-x