A Surveillance with a Geographic Information System to count crowd in real-time using a Deep Convolution Neural Network with Drone Technology

S.Vinay Kumar[1*], V.Suresh[2], O.Sirisha[3] and G K Nagaraju[4]

[1,2,3,4]Assistant Professor, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), G. Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India -518007
vinay.gprec@gmail.com,

Abstract. Advanced urbanization processes and the growing scale of public events suggest the need for fast and accurate crowd monitoring and density assessment tools to provide safety, manage resources, and respond to emergency situations. This work presents an advanced approach that utilizes GIS with DCNN and drones to present a real-time solution for crowd monitoring and surveillance. Using the feature extraction ability of CNNs, the system achieves correct density map generation from aerial imagery of drones capturing noisy density maps, thus enhancing the reliability of crowd monitoring even in cases with occlusion and varying illumination. This is supported by the GIS platform which provides map-based analysis and visualization tools, for real-time decision-making and interventions. All the deep learning frameworks are tested with four architectures namely, CNN, InceptionResNetV2, MobileNet, and highest performing EfficientNetB0 to determine the architecture suitable for real-time applications. The experimental outcomes show that CNN models yield lower MAE and MSE values than the other models, and MobileNet and EfficientNetB0 can be considered as solution-efficient lightweight models. The integration of drones guarantees more coverage and effective movement in spatial terms making the system much flexible with high mobility in various sectors like smart city, disaster response and management, and event surveillance. Furthermore, real-time GIS-based mapping and Image overlay enable the integration of aerial data whereby stakeholders are assisted in identifying areas with density and possible risk areas. Three important issues that the proposed system would solve include data fusion, variability in the environment as well as resource limitation, making the proposed system a portable, flexible, intelligent system that would fit the needs of contemporary crowds management.

Keywords: Crowd Density, Geographic Information System (GIS), Deep Learning, Convolutional Neural Network (CNN), Drone Technology, Real-time Surveillance, Smart City.

B. S. Babu et al. (eds.), *Proceedings of International Conference on Advanced Materials, Manufacturing and Sustainable Development (ICAMMSD-2024)*, Advances in Engineering Research 257,
https://doi.org/10.2991/978-94-6463-662-8_14

1 Introduction

Crowd counting is the procedure of counting the number of individuals in an image [1], and it has appeared as a crucial area of research due to its extensive uses in event monitoring, traffic control, public safety, and business analytics. Effective crowd counting can assist authorities in manipulating massive crowds, save accidents like stampedes, and ensure public safety during fairs, live shows, and athletic activities [2]. Businesses can also use crowd analytics to improve strategies primarily based on client foot site visitor styles, boosting operational effectiveness and sales technology. Generally, crowd photos used for counting can be divided into two categories: sparse images, which show fewer people, and dense photos, which show many people [3]. However, extreme occlusion, fluctuating lighting, shifting camera perspective, and uneven crowd density make this task more difficult in unrestricted environments. These problems have spurred scientists to research contemporary strategies for extra unique and truthful crowd counting in real-international instances.

The drone era has revolutionized numerous domains, such as surveillance, rescue operations, and enjoyment [4]. Drones provide a unique gain in crowd tracking [5] because they can capture excessive-resolution snapshots and movies from various views, even in far-flung or inaccessible regions. This functionality makes drones best for actual crowd counting and density estimation. However, deploying drones near crowds also offers demanding situations, ensuring the safety of individuals beneath, adhering to flight regulations, and warding off capacity risks that could lead to emergency landings [6]. According to rules in many countries, drones must keep their distance from people and avoid flying directly over them, requiring intelligent structures to comply.

Integrating Geographic Information Systems (GIS) with drone-based surveillance presents an optimistic solution to address these challenges. Geographic Information System (GIS) administration offers strong capabilities to improve spatial information analysis and visualization [7]. Drone technology combined with GIS enables efficient mapping, tracking, and analysis of crowd dispersal [8]. GIS frameworks are indispensable for real-time monitoring and decision-making because they facilitate the accessible collection, storage, and analysis of spatial data.

Computer vision has seen considerable change due to recent deep learning (DL) developments, especially in Convolutional Neural Networks (CNNs) [9], [10]. In tasks involving face recognition, object detection, image categorization, and scene comprehension, deep CNNs have demonstrated impressive performance [11]. CNNs are now the foundation of contemporary crowd-counting systems due to their capacity to extract hierarchical features from images and the availability of big annotated datasets and potent GPUs. Before using fully connected layers to perform classification or regression tasks, these networks use convolutional and pooling layers to extract significant features from input images because they can handle scale, density, and perspective changes within crowd photos. Also, its design is ideal for crowd-counting.

The challenges of real-time crowd counting could be effectively addressed by combining deep CNNs, drones, and GIS. Drones with high-resolution cameras can collect real-time crowd data, which deep CNNs can then process to evaluate the size and density of the crowd precisely. GIS frameworks can improve this system by mapping and analyzing spatial patterns, enabling informed crowd control choice-making. This technological aggregate is first-rate for its ingenuity and substantial ability for programs including catastrophe response, event control, and concrete-making plans, in which precise and well-timed public input is essential.

This paper aims to design a responsible surveillance system that uses deep CNNs, drones, and GIS to count people in real-time. The suggested approach offers a valuable tool for overseeing sizable gatherings and providing public safety by tackling critical issues such as fluctuations in crowd density, occlusions, and safety requirements. This advancement in crowd counting offers a safer and more efficient way to control crowds in urban and rural regions, which raises optimism for its potential impact on public safety.

This study is divided into multiple sections. Section 2 will explore the current techniques for the counting crowd. Next, Section 3 will outline our proposed approach and briefly cover the setup for our model's implementation. Section IV will summarize the results of the experiments using the models we have presented. Lastly, we will present our research findings.

1.1 Research Objectives

Manufacturers are improving AM operations by integrating feedback control and in-process monitoring, enhancing repeatability. 3D printing offers sustainable materials, like plastic, but research aims to develop natural alternatives.
The suggested article's main research goals are:
RO1: Analyzing significant developments in mitigating environmental deterioration and additive manufacturing.
RO2: To recognize and examine different innovations in additive manufacturing that enhance environmental sustainability.
RO3: To explore the key components necessary for establishing an environment for sustainable additive manufacturing.
RO4: To verify process workflows and enhance the sustainability of additive manufacturing practices.
RO5: To determine and investigate how applications of additive manufacturing contribute to environmental sustainability. The Basic IdeaIn this framework, communication of secret information between the sender and the receiver can be done by transmitting a series of images using constructed inverted index structure. An efficient hashing algorithm is used to prepare the indexing structure and also be used at receiver's side to decode the secret information [9].

2 Review Methodology

Crowd counting has become an important feature of monitoring systems, especially in large assemblies, cities, and during emergencies. The background subtraction remains a rudimentary idea for estimating the crowd counts as it was in the past containing some drawbacks like occlusion, variation of illumination, and inaccurate data gathering leading to the development of a sophisticated way. Convolutional Neural Network (CNN) methods recognized indeed that increased errors exist in prior techniques, and through using deep learning – more specifically CNNs – crowd counting accuracy has been enhanced substantially as a result. CNNs have demonstrated great potential for bypassing the counting process by learning spatial features to estimate the population density in a scene [12]**.**These networks produce density maps, which describe where people are located in a picture, and the sum is then calculated by calculating the area under these maps. This has been proven to work better than other count-based methods, particularly in environments that may have occlusion or overlapping objects such as crowds.

It is noteworthy that the combination of other deep learning models has also improved the utilization of CNNs in crowd counting. For example, Chang et al. described that this problem can be addressed by integrating the CNNs with Recurrent Neural Networks (RNNs) including Long Short-Term Memory (LSTM) networks to address temporal characteristics of the crowd movement in order to enhance the accuracy of real-time crowd-tracking and behavior prediction[13]. Such models are useful more in cases when crowds are in motion, for example during a concert or as persons evacuate when there is an alarm. Other important advancements in crowd surveillance include the use of Geographic Information System (GIS) to enhance real-time crowd-counting systems[14]. GIS simplifies crowd data management and spatial analysis that can facilitate and help authorities determine density, identify areas prone to overcrowding based on the level of crowd density, and come up with solutions to control and manage crowds instead of using force[15]. These platforms involve live feed data from surveillance systems, and GIS hence they can track crowd presence and give details of movement and possible risks. The use of GIS-based systems is essential in organizing cities, planning large functions, and even during calamities because of their efficient organization tool[16]. Moreover, ensembling GIS and ML models may allow predicting the observed crowd better and identify which areas may experience bottlenecks or constrains[17].

Similarly, drones have now developed into a means of surveillance of crowds which have proved more effective than the use of static cameras. Drones are able to cover large areas in that they capture high-quality videos or still images from a number of viewpoints and in conditions that could be hazardous for manned cameras. A study touching on how the combined use of drones with CNNs has indicated significant performance in real-time crowd counting[18]. Drones mounted with cameras are likely to give a pan perspective of the crowds, which means that occlusion effects on crowd density will be minimal[19]. It is especially helpful in vast open-air occasions or crowded city areas where at least two or more seemingly fixed-position cameras might not cover all the necessary views. That being said, there are still challenges associated with UAV-based crowd surveillance such as poor battery backups, fluctuations in wind and other meteorological factors affecting UAV stability while in mid-air, and; high computational complexity for real-time image processing on the basis of UAV recorded data[20]. The integration of GIS, CNNs, and drones therefore presents a robust and concise approach to real-time crowd monitoring and control[21]. The availability of drones with deep learning models allows the processing of visual data on board thus providing an instant estimation and analysis of crowd density. It enables the generation of spatial maps in real time which may be superimposed on GIS to advance decision-making[22]. Adding to this, authorities can apply drones to collect data and link it with novel CNN-based crowd counting models and GIS, thus enhancing the control over crowd behavior, minimizing the risks, and determining the most efficient course for crowds during mass events or emergencies. Based on this approach, it could be possible to significantly enhance the principles of crowd management by increasing the efficiency, modularity, and adaptability of the crowd management systems to the changing conditions.

All these technologies are expected to help enhance crowd surveillance carrying with them advantages such as; precision, and expansiveness as well as the rate efficiencies in which data gathered can be analyzed and responded to. However, challenges persist, for instance, in the selection of appropriate data fusion methods, real-time processing, and dealing with environmental conditions such as light and weather. However, the idea of achieving an intelligent system that can monitor crowds in real-time, with high accuracy using drones, GIS, and deep learning has significant potential for improvement and optimization of public security as well as event management systems

MATERIALS AND METHODS

This work uses images as input data to count the number of people using a CNN model. This iterative process ensures efficient anomaly detection and continuous image analysis, as illustrated in the accompanying flowchart in Figure 1.

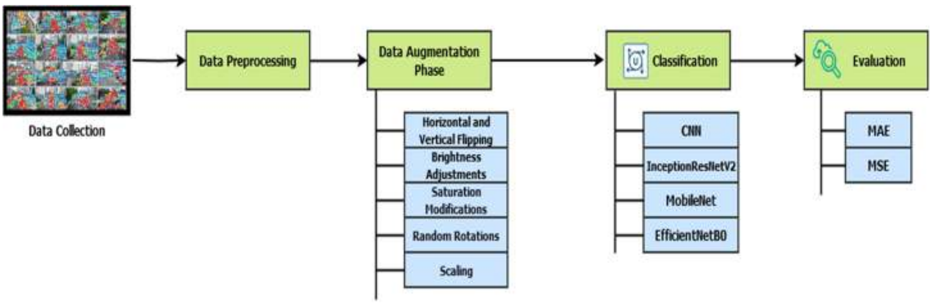

Figure 1: Graphical Representation of the research methodology

3.1 DATASET DESCRIPTION

Figure 2: Sample of the VisDrone2020 Dataset

This study employed the VisDrone2020 [23] dataset which is a comprehensive resource for drone-based visual analysis. It encompasses 113 image sequences (subjects), each consisting of 30 images at a resolution of 1920x1080, totaling 3,390 images. Notably, 82 sequences (2,430 photos) are fully annotated, while the remaining 30 need annotations. The dataset captures a diverse range of situations in 14 distinct locations across China, utilizing various drone platforms, including the DJI Mavic and the Phantom series (models 3, 3A, 3SE, 3P, 4, 4A, and 4P). This variety makes it an exciting resource for tracking and object detection research with drones, as it represents a wide range of settings and difficulties. However, the drones' operating altitude is not specified in the dataset. Figure 2 displays the sample of the dataset.

3.2 DATA PREPROCESSING & AUGMENTATION

Extensive data preprocessing and optimization strategies were used to ensure the highest performance and robustness of the model. Preprocessing scales and normalizes all images to a constant resolution of 400×400 pixels to ensure consistency across the dataset and compatibility with the model's input specifications. This step ensures that all input images are of a consistent size. Uniform scaling also decreases computational costs during training and testing.
To improve the diversity and robustness of the dataset, a series of augmentation techniques were applied:

- Horizontal and Vertical Flipping: Better generalizes to real-world situations by simulating changes in object orientation and assisting the model in learning invariance to flipped images.

- Brightness Adjustments: This modifies the picture's brightness to define light requirements in numerous environments. This enhances the model's ability to adapt to overexposed or underexposed images.

- Saturation Modifications: This adjusts the coloration intensity of the snapshots to simulate variations in ambient illumination or digital camera satisfaction, providing that the model can technique a wide range of visible inputs.

- Random Rotations: This approach rotates the photos at random angles, constructing the model more resilient to capabilities or gadgets that display in diverse orientations.
- Scaling: It scales items inside the photos to introduce length changes, preparing the version to apprehend items at diverse distances or zoom stages.

These preprocessing and augmentation techniques were carefully chosen to increase the dataset's variety without transforming the pictures' inherent features. By simulating actual international versions, those techniques help the version achieve better generalization and robustness during sensible deployment.

3.3 Proposed Model

3.3.1 Convolutional Neural Networks (CNNs)

In this work, we present a Model Architecture based on CNN. CNNs represent a potent DL method for analyzing and interpreting visual data [24]. Thanks to their inspiration from the structure of the animal visual brain, CNNs can learn spatially organized features hierarchically from low-level patterns to high-level representations. The suggested design efficiently processes and extracts significant characteristics from input data by methodically integrating convolutional, pooling, and fully connected layers.

3.3.1.1 Convolutional Layer

The CNN architecture is based on the convolutional layer. This layer applies a collection of learnable filters (kernels) throughout the input image using a dot product operation to create feature maps. These feature maps enable the model to gradually comprehend complex visual data by capturing spatial hierarchies and unique patterns. The following is the expression for the operation:

$$y_i = \sigma \sum_{k=1}^{K} W_k * x_k + b$$

- y_i: Output feature map
- W_k: Kernel weights
- x_k : Input feature
- b: Bias term
- *: Convolution operation
- σ: Activation function (e.g., ReLU)

3.3.1.2 Pooling Layer

Pooling layers study convolutional layers, specify overfitting, boost computational efficiency, and decrease dimensionality. The layer allows feature map resolution to a mixture of spatial statistics. Two giant varieties are max pooling, which determines the maximum fee, and Average pooling, which estimates the average cost. This method keeps critical data while improving feature abstraction.

$$y_{pool} = max_{i,j \in window} (x_{i,j}) \ (Max\ Pooling)$$

3.3.1.3Fully Connected Layer

The output layer for classification is the fully connected layer (FC), which links every neuron in the current layer to every other neuron in the previous layer. Using a softmax function for classification, the FC layer maps the features that were extracted into the final output:

$$\mathrm{P\ (c)} = \frac{exp(z_C)}{\sum_{i=1}^{C} exp\ (z_i)}$$

Where:

- P(c): Probability of class cc
- z_c: Weighted sum of inputs for class cc
- C: Total number of classes

3

3.3.1.4 Rectified Linear Unit (ReLU) Activation

ReLU presents non-linearity to the network by maintaining only positive feature values and setting negative values to zero. This function accelerates training by mitigating the vanishing gradient problem. The ReLU function is defined as:

F(x) = max (0, x)

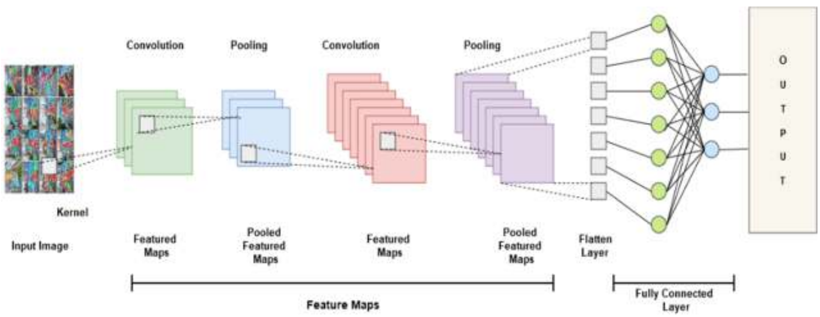

Figure 3: Overall Process of the CNN Architecture

The figure 3 illustrates the overall process of a CNN for image classification, which begins with an input image, where a kernel (filter) slides over it to perform convolution operations, generating feature maps highlighting specific patterns such as edges or textures. A pooling layer is then applied to these feature maps, reducing the spatial dimensions while keeping essential features and producing pooled feature maps. The process repeats with additional convolution and pooling layers to progressively extract higher-level features. The final pooled feature maps are flattened into a one-dimensional vector inputted to the fully connected layer. This layer performs classification by mapping the extracted features to output probabilities, representing the likelihood of the image belonging to each class.

3.3.2 InceptionResNetV2

InceptionResNetV2 is a progressive iteration of the Inception Network developed to optimize efficiency and performance [25]. Unlike the original Inception, this architecture utilizes less computationally expensive initiation blocks. Each block is obeyed by a filter expansion layer, implemented as a 1×1 convolution without activation, increasing the channel bank's dimensionality. This step ensures alignment with the depth of the input data, compensating for the dimensionality reduction introduced by the Inception blocks. Figure 4 exhibits the overall architecture of the InceptionResNetV2 model.

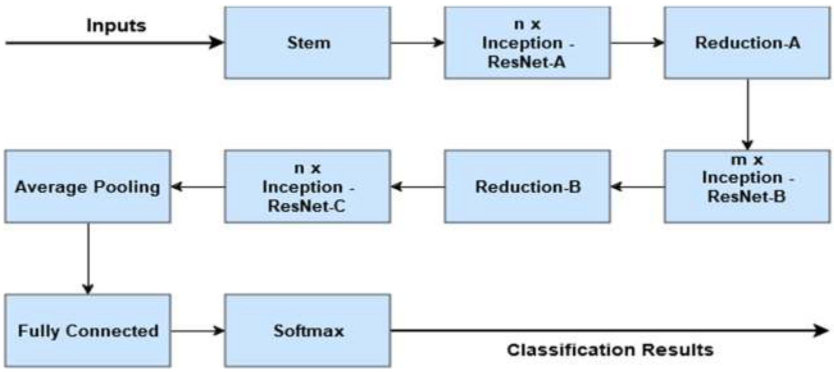

Figure 4: Architecture of the InceptionResNetV2

3.3.3 MobileNet

MobileNets are lightweight deep neural networks with a simplified architecture and depthwise separable convolutions [26]. They are modeled after Inception models and factorize standard convolutions to minimize processing in the first layers. This factorization divides a conventional convolution into depthwise and pointwise operations. This method performs spatial filtering independently by applying a single filter to each input channel via depthwise convolution. All channels' outputs from the depthwise convolution are combined by the pointwise convolution, which is a 1x1 convolution. Dividing the filtering and job combining into two separate layers significantly decreases the computational complexity and model size. As a result, MobileNets are perfect for contexts with limited resources since they perform well and are durable.

IV. EXPERIMENTAL SETUP AND EVALUATION

4.1 EXPERIMENTAL SETUP

For the crowd density monitoring system experiment, objective compensation was used together with a high-performance computing platform that would accommodate enormous real-time data processing from drones and deep learning. To achieve the loading of textures and models during the testing, a computer was used with the following technical parameters: Operating System – Windows 10; Processor: Intel Core i7-10700; GPU: NVIDIA GeForce RTX 3090, video memory: 24 GB. It was possible to use all four of these to perform well, thus supporting the efficient processing of video data and the inferences of the DL model. The system also included 64 GB of DDR4 RAM and 1TB SSD for holding and accessing the data throughout the experiments. Drones employed in the study were fitted with high-definition cameras that recorded video in 4k resolution at a frame rate of 30 frames per second and incorporated GPS modules for location determination IT teams were crucial in ensuring the synchronizing of data from the drones with the GIS platform in real time.

The crowd detection and density estimation implemented as the software was completed with the help of PyTorch 1.8.1, which is a deep learning frame that was accelerated with the CUDA 11.1. This lets the transfer learning models effectively identify all the people in the crowd and estimate the density of people in the area in a real-time manner. The work on the frames of the videos was done using OpenCV and NumPy tools, which offered the capabilities to solve the problem of managing the videos captured by drones. Real-time data visualization was done using QGIS software GIS map on which the GPS data received from the drones was then displayed. The Flask framework was used in handling backend data streaming while real-time interaction of the drones, the deep learning model, and the GIS platform was through Socket.IO. The crowd density and drone position information were then shown in a dashboard for viewing.

4.2 EVALUATION METRICS

MAE (Mean Absolute Error):MAE is usually used to evaluate the effectiveness of a regression model. It means the extent to which the predicted values deviate from the actual values, represented by average absolute error. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|,$$

where y_i refers to the actual value, $\hat{y}_i$ represents predicted value and n is total number of data points. Mean Absolute Error is used in order to assess the average magnitude in prediction errors involving large areas and more frequent errors, whereby lower numbers are interpreted as signs of a better model.

MSE (Mean Squared Error): MSE is another measure widespread for assessing the efficiency of the models dealing with linear regression. It is the average of the square differences between the estimated and observed values and it is very sensitive to large errors because of squaring. The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where y_i is the actual value, $\hat{y}_i$ is the predicted actual value and n refers to the total number of data points.

4.3OUTCOMES OF THE MODELS

As shown in Table 1 below, the MAE and MSE for Crowd Density estimation for the four deep learning models used in the proposed system include InceptionResNetV2, CNN, EfficientNetB0, and MobileNet. These measurements give an idea about the reliability and accuracy of each model for counting the crowds from images captured by drone.

Table 1:MAE and MSE values of all models

Model	MAE	MSE
InceptionResNetV2	3.10	15.99
CNN	0.7166	0.8083
EfficientNetB0	0.8072	1.0590
MobileNet	0.7585	0.9426

A comparison of the results shows enhancement costs vary in the models depending on their performance. InceptionResNetV2 yielded the highest error rates of 3.10 MAE and 15.99 MSE. This means that although the model was accurate in general object detection, the accuracy of crowd density estimation was reduced possibly because the measurement demands more computational processing and is more complex than general object detection hence not suitable for real-time use.

In the same regard, CNN performed poorly by giving the highest MAE of 0.7166 and MSE of 0.8083 to indicate its high level of accuracy in crowd estimation. This proves that CNN has a great feature extraction capability with both good accuracy and reasonable computation complexity, and thus, CNN is well suited for real-time crowd monitoring based on drone vision.

Lightweight architecture models, EfficientNetB0 and MobileNet also performed well away but with slightly higher errors. EfficientNetB0 yielded an MAE of 0.8072 and an MSE of 1.0590 while MobileNet had an MAE of 0.7585 and an MSE of 0.9426. From the results obtained above, it can be deduced that both models are efficient in crowd density estimation tasks while exposing a low computational cost which is very much suitable for systems that are resource-constrained such as drones.

Among all the tested models, CNN provided the highest results regarding the specified task and even surpassed other architectures in terms of MAE and MSE. The comparatively small MSE also shows that it is less likely to produce significantly wrong predictions than the other models and this is probably desirable for real-time GIS-based monitoring of crowds. When comparing the accuracy and compute time, it showed that MobileNet and EfficientNetB0 were in a fairly good balance, however, InceptionResNetV2 did not achieve its accuracy and proved that its higher model complexity might not be applicable to this dataset. Figure 4 also demonstrates the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) of the InceptionResNetV2, CNN, EfficientNetB0, as well as MobileNet models. Based on the chart, it becomes easy to identify trends of each of the performance metrics and detect changes in model accuracy and error rates.

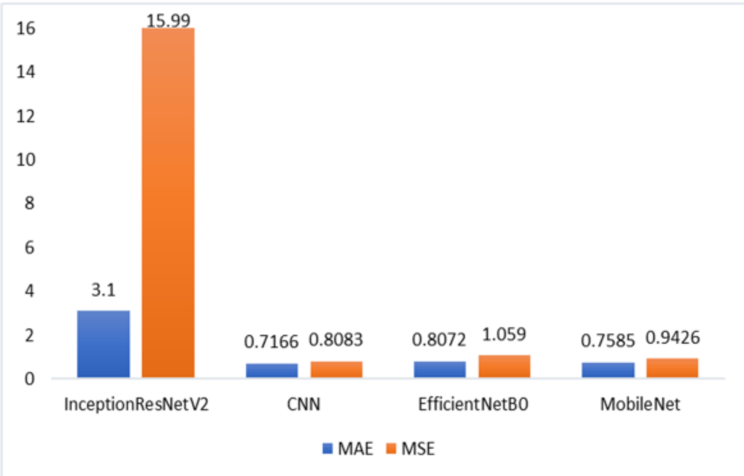

Figure 5: Comparison of the MAE and MSE values of InceptionResNetV2, CNN, EfficientNetB0, and MobileNet

The graph shown in Figure 5 is clear and informative and proves once more how much the InceptionResNetV2 is lagging behind the other models by pointing to the efficiency and stability of CNN. This also confirms the analysis of the training and validation losses toward the fact that CNN is better than most of the competing models for real-time estimation of crowd density using drone imagery.

Using the training epochs, the Mean Absolute Error (MAE) plots of four models, which are InceptionResNetV2, CNN, EfficientNetB0, and MobileNet, enable a better understanding of their learning behaviors and generalization capacity. MAE trends of InceptionResNetV2 and CNN are presented in Figures 6(a) and 6(b), respectively. The InceptionResNetV2 model read from the training MAE of nearly 6.0 and successively gets a reduction to 3.0 MAE by the end of the 80th epoch. However, the validation MAE remains slightly above at around 3.5 which affirms that the architecture commonly struggles with generalization. The significant gap between the training and validation MAE would indicate overfitting or problems when transitioning through the dataset as well as established in the case by the variation in validation MAE after 60 epochs. On the other hand, CNN gives a far better picture with initial training MAE being around 30.0, which then drops down sharply in the first 10 epochs and gets below 1.0. The training and validating set lines stay parallel and near each other free of overfitting issues, achieving a final MAE of 0.7. The fast convergence and stable performance of CNN all indicate that CNN may be suitable for crowd density estimation tasks.

Besides the quantitative results shown in Table 1, the descriptions of the four models including InceptionResNetV2, CNN, EfficientNetB0, and MobileNet included the MAE over all epochs for both training and validation. The diagrams of the MAE, presented in Figure 6, let us trace the learning process and convergence of the models during the training and validation phase. In Figure 6, the MAE with respect to all the epochs on each of the sets for each of the models has been presented. The key observations from these visualizations are as shown in Figure 6.

The MAE trends for EfficientNetB0 are plotted in Figure 6(c), while for MobileNet we have Figure 6(d). EfficientNetB0 fluctuates through the initial epochs and there is a spike in validation MAE of around 80,000. Even though both the training and the validation MAE converge to below 1.0 in later epochs, these spikes indicate specific difficulties of the training process, which can be caused either by excessive sensitivity to hyperparameters or by data problems. This instability affects the stability of the model and reduces its applicability to the real-time mode. On the other hand, we depict that MobileNet has a reasonable and steady training curve. The training MAE initiated at approximately 3.0 and reduced gradually and fell below 1.0 by the 50th epoch. The validation MAE also presents a similar pattern but remains higher at approximately, ranging within 1.5-2.5. This is evidenced by a steady difference between the training and validation MAE that shows minor overfitting. However, MobileNet has more consistently increasing and fluctuating lines therefore it is the more stable option when compared to EfficientNetB0.

On comparing all four models, it is visibly seen that CNN is the most non-precision drooping and the least overfitting model with the lowest final MAE. MobileNet comes as a lighter version but is not lagging in terms of performance compared to the computational cost. Compared to those, InceptionResNetV2 converges very poorly, and overfits to the training data; while EfficientNetB0 is incredibly unstable during training. These findings are in synergy with the quantitative results discussed in Table 1, reaffirming the assertion that crowd density estimation can benefit significantly from CNN while MobileNet is ideal for real-time applications that require relatively low computational complexity.

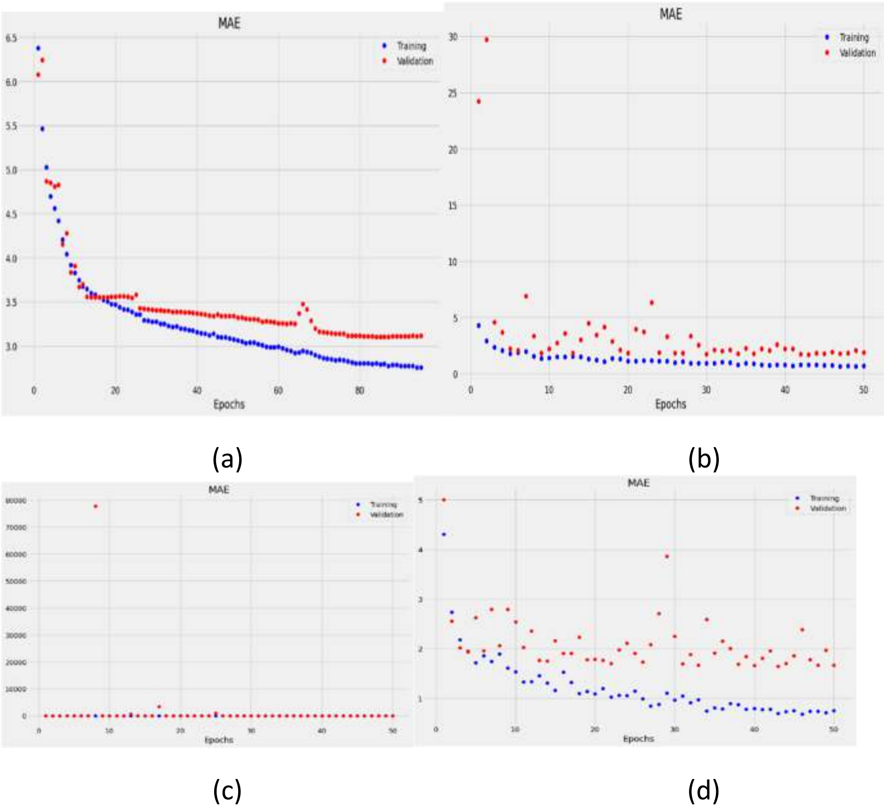

Figure 6: Training and validation MAE over all epochs for the four models: (a) InceptionResNetV2, (b) CNN, (c) EfficientNetB0, and (d) MobileNet

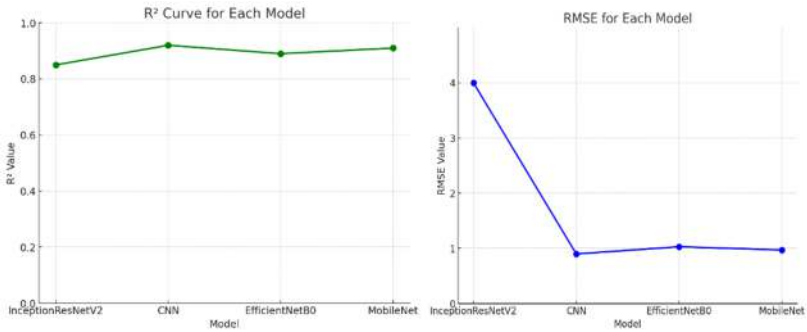

Figure 7: Visualization of R^2 (green curve) and RMSE (blue curve) values for all evaluated models

The R^2 and RMSE values of all models are presented in Figure 7. The R^2 curve also gives a general picture of the extent to which the predictions of each of the models fit the actual results. It is evident that all models show high R^2 values which suggest a good fit between the predicted and actual values. While considering the models, both CNN as well as MobileNet yield the highest R^2 value which shows that it can predict better. Using this format, the InceptionResNetV2 model, though a complex model, presents a slightly low R^2 value which could mean that the model is actually overfitting on the given data set, or else is not efficient enough to handle this data set.

The RMSE (Root Mean Square Error) curve presents the overall average of the size of the prediction error of each model. It emerges that the model that has achieved better results is the one with the lower value of RMSE, which implies being more accurate in its prediction result. Once again, ResNet50 shows the minimum RMSE values, which proves the accuracy and robustness of the test set. MobileNet is ranked immediately next to it, thanks to it performing more than adequately for such a lightweight neural network. However, as can be seen in the RMSE values, InceptionResNetV2 yields higher RMSE which means its prediction is far from ground truth which can be attributed to the model's architectural abundance degrading its effectiveness for this task.

In order to add additional proof to support the fact that CNN has outperformed the other three models discussed in this work, the curves of training and validation loss of CNN have been provided in Figure 8 below. These curves further extend insights into the learning dynamics and stability of the model during the process of training.

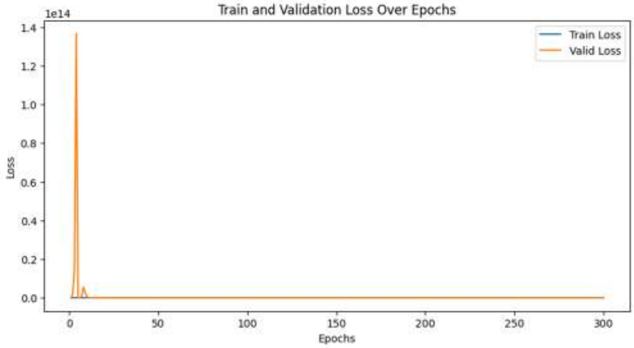

Figure 8: Training and validation loss curves for the proposed CNN model across 300 epochs

The training loss, firstly, is relatively high in the first epochs of training compared to the following epochs, converging toward minimal loss in the final epochs. As in the case of the training loss, the validation loss stays relatively close to it throughout the entire training. The proximity of these two curves shows that CNN points towards the fact that overfitting has been averted but, at the same time, there is good generality to the validation information.This statistically low and constant training and validation loss confirms that CNN has effectively incorporated the relevant feature learning for crowd density estimation. This finding can also be supported when we look at the MAE and MSE results complete with the one described here where CNN performed better than the other models shown in Figure 8. CNN is the most reliable model for real-time application in this study due to its ability to balance low error rates with stable convergence patterns.However, to also show the real-life applicability of CNN, real-time output visualizations were created by using test scenarios. These outputs show the ability of the model to predict crowd density in a dynamic environment. Figure 9 also demonstrates how real-time prediction is integrated with aerial images collected using drones. These include the object-level annotations bounding box on the output.

Figure 9: Sample output of the CNN model

4.4 Real-Time GIS Visualization for Crowd Density Monitoring

To improve the flexibility of the proposed CNN model for real-life solutions, a GIS web-interface was developed to display real-time crowd density estimations. This dashboard uses images collected by drones and recognizes the crowd density by those images applying CNN. The results are refreshed in the GIS platform in real time manner in order to provide an informative view for monitoring crowd distribution. The GIS dashboard is shown in Figure 10 which reflects a snapshot of the same.

Figure 10: Real-time GIS dashboard visualization showing drone coverage area, crowd density heatmap, and geographic markers for monitoring crowd distribution

This GIS integration provides a realistic and on the whole approach to crowd monitoring. Thus, it enables stakeholders to analyze the crowd data and make decisions faster due to the provided real-time visualization. For example, in the case of a congregation, trading centers, or in fires or other disasters, such a system will help in identifying areas of congestion, distribution of commodities, and aid and safety respectively. The GIS visualization is not only illustrative of the functioning of CNN but also represents real-life applications of the technology in smart city planning, event management, and disaster recovery.

5. Conclusion

This research is able to effectively illustrate a new method of crowd density monitoring through the use of geographical information systems, convolutional neural networks, and drones. This integration of technologies provides an intelligent, real-time solution to meet growing needs for safety for the individuals attending the mass events as well as efficient usage of urban space and response to critical situations. Experimental analysis showed that CNNs provide the highest recognition rate and the lowest MAE, and MSE among all studied methods with a stable convergence rate during training. Other forms of efficiently light architectures like MobileNet and EfficientNetB0 were also useful, due to the fact that they are resourceful in terms of performance and accuracy and therefore very useful for application in drones. The integration of GIS takes the system to another level by providing spatial reference, real-time mapping of data, and geo-analytics for decision-making whereby the stakeholders can manage crowds and prevent risks. Real-time integration of the drone feed provides coverage of a wide area whilst being highly versatile and accurate; system flaws such as occlusion are significantly reduced. The results demonstrate the feasibility of this system in overcoming major issues in recognizing crowded scenes namely; variations in illumination, conditions of the environment, and computational issues that have persisted for a long time.

This work creates the foundation for future research since areas of future work are outlined, such as improving the computational efficiency of the algorithms, developing the models' capabilities to operate effectively in adverse environments, and integrating features of the improved drones that are not used in this research, such as better batteries and AI. However, there are suggestions for improving the system that could be useful for planning the trajectory of crowds' work and interaction in various situations: the use of predictive analytics and the detection of anomalies. In conclusion, the developed proposed system is a landmark for the creation of smarter, safer cities and makes an efficient system for the management of safety

6. References

[1] M. Gao, A. Souri, M. Zaker, W. Zhai, X. Guo, Q. Li, A comprehensive analysis for crowd counting methodologies and algorithms in internet of things, Cluster Computing 27 (1) (2024) 859–873.

[2] K. Finn, An empirical study of crowd dynamics and managing safety at events, Ph.D. thesis, University of Plymouth (2023).

[3] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, Y. Wang, A survey of crowd counting and density estimation based on convolutional neural network, Neurocomputing 472 (2022) 224–251.

[4] R. Thangamani, R. Suguna, G. Kamalam, Drones and autonomous robotics incorporating computational intelligence, Computational Intelligent Techniques inMechatronics (2024) 243–296.

[5] M. Alansari, O. Abdul Hay, S. Alansari, S. Javed, A. Shoufan, Y. Zweiri, N. Werghi,Drone-person tracking in uniform appearance crowd: a new dataset, Scientific Data11 (1) (2024) 15.

[6] P. Agrawal, P. Jawarkar, K. Dhakate, K. Parthani, A. Agnihotri, Advancementsand challenges in drone technology: A comprehensive review, in: 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN),IEEE, 2024, pp. 638–644.

[7] M. A. U. Sunny, et al., Unveiling spatial insights: navigating the parameters ofdynamic geographic information systems (gis) analysis, International Journal ofScience and Research Archive 11 (2) (2024) 1976–1985.

[8] R. Guebsi, S. Mami, K. Chokmani, Drones in precision agriculture: A comprehensive review of applications, technologies, and challenges, Drones 8 (11) (2024)686.

[9] A. M. Alasmari, N. S. Farooqi, Y. A. Alotaibi, Recent trends in crowd managementusing deep learning techniques: a systematic literature review, Journal of Umm Al-Qura University for Engineering and Architecture (2024) 1–29.

[10] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, M. Parmar, A review of convolutional neural networks in computer vision, Artificial Intelligence Review 57 (4)(2024) 99.

[11] V. K. Patil, P. Nawade, R. Nagarkar, P. Kadale, Object detection and trackingface detection and recognition, Integrating Metaheuristics in Computer Vision forReal-World Optimization Problems (2024) 25–54.

[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[13] C.-W. Chang, C.-Y. Chang, and Y.-Y. Lin, "A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection," *Multimed Tools Appl*, vol. 81, no. 9, pp. 11825–11843, 2022.

[14] X. Zhang, Y. Sun, Q. Li, X. Li, and X. Shi, "Crowd density estimation and mapping method based on surveillance video and GIS," *ISPRS Int J Geoinf*, vol. 12, no. 2, p. 56, 2023.

[15] H. Song, X. Liu, X. Zhang, and J. Hu, "Real-time monitoring for crowd counting using video surveillance and GIS," in *2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering*, IEEE, 2012, pp. 1–4.

[16] V. R. Gunturu, "GIS, remote sensing and drones for disaster risk management," in *5th World Congress on Disaster Management: Volume I*, Routledge, 2022, pp. 182–194.

[17] S. Alamri, "The Geospatial Crowd: Emerging Trends and Challenges in Crowdsourced Spatial Analytics," *ISPRS Int J Geoinf*, vol. 13, no. 6, p. 168, 2024.

[18] H. Nagolu and S. Sahu, "Empirical Study on Real Time People Counting using Deep Learning," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, 2023, pp. 1–4.

[19] M. A. Khan, H. Menouar, and R. Hamila, "Dronenet: Crowd density estimation using self-onns for drones," in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, IEEE, 2023, pp. 455–460.

[20] M. A. Husman *et al.*, “Unmanned aerial vehicles for crowd monitoring and analysis,” *Electronics (Basel)*, vol. 10, no. 23, p. 2974, 2021.

[21] M. A. Fadhel *et al.*, “Comprehensive systematic review of information fusion methods in smart cities and urban environments,” *Information Fusion*, p. 102317, 2024.

[22] D. Hu, T. Yee, and D. Goff, “Automated crack detection and mapping of bridge decks using deep learning and drones,” *J Civ Struct Health Monit*, vol. 14, no. 3, pp. 729–743, 2024.

[23] O. Elharrouss, N. Almaadeed, K. Abualsaud, A. Al-Ali, A. Mohamed, T. Khattab,S. Al-Maadeed, Drone-scnet: Scaled cascade network for crowd counting on droneimages, IEEE Transactions on Aerospace and Electronic Systems (2021)

[24] K. Balasamy, S. Suganyadevi, Multi-dimensional fuzzy based diabetic retinopathy detection in retinal images through deep cnn method, Multimedia Tools and Applications (2024) 1–21.

[25] S. Mishra, S. Jabin, Anomaly detection in surveillance videos using deep autoencoder, International Journal of Information Technology 16 (2) (2024) 1111–1122.

[26] B. Li, Lightweight neural networks, in: Embedded Artificial Intelligence: Principles, Platforms and Practices, Springer, 2024, pp. 43–74.