PeopleNet: A Novel People Counting Framework for Head-Mounted Moving Camera Videos

Ankit Tomar, Santosh Kumar, Bhasker Pant

Graphic Era Deemed To Be University, Dehradun (India)

Received 26 Mayo 2022 | Accepted 2 February 2023 | Early Access 12 April 2023

Abstract

Traditional crowd counting (optical flow or feature matching) techniques have been upgraded to deep learning (DL) models due to their lack of automatic feature extraction and low-precision outcomes. Most of these models were tested on surveillance scene crowd datasets captured by stationary shooting equipment. It is very challenging to perform people counting from the videos shot with a head-mounted moving camera; this is mainly due to mixing the temporal information of the moving crowd with the induced camera motion. This study proposed a transfer learning-based PeopleNet model to tackle this significant problem. For this, we have made some significant changes to the standard VGG16 model, by disabling top convolutional blocks and replacing its standard fully connected layers with some new fully connected and dense layers. The strong transfer learning capability of the VGG16 network yields in-depth insights of the PeopleNet into the good quality of density maps resulting in highly accurate crowd estimation. The performance of the proposed model has been tested over a self-generated image database prepared from moving camera video clips, as there is no public and benchmark dataset for this work. The proposed framework has given promising results on various crowd categories such as dense, sparse, average, etc. To ensure versatility, we have done self and cross-evaluation on various crowd counting models and datasets, which proves the importance of the PeopleNet model in adverse defense of society.

Keywords

Deep Learning, Density Map, Feature of Expansion, Moving Camera Videos, People Counting.

DOI: 10.9781/ijimai.2023.04.002

I. Introduction

Object detection and counting are emerging issues for the development of social sectors such as agriculture, wildlife sustainability, satellite imaging, drug molecule detection, crowd protection, etc. Computer vision (CV) facilitates pedestrian estimation to solve the social and administrative congestion monitoring problems which is a burning issue nowadays [1]. Automatic crowd counting is easier in video surveillance when the imager is stationary; however, it is more challenging in videos consisting of background motion driven by a moving camera and moving objects. Most crowd image databases are created by capturing from stationary cameras, resulting in a large number of occlusive samples, which limits the performance of any crowd counting (CC) mechanism. The existence of moving background, size of people, motion vibration, and camera position are some of the natural obstacles in crowd samples that also limit the people counting performance. There are some universally agreed challenges faced while developing an automated CC framework. Providing fair distribution of training information over live video streaming is the most common challenge [2] to protect the privacy of any individual by intentionally or unintentionally targeting individuals. Establishing a fair relationship between society and the monitoring system is another challenge to developing a reliable CC system. In addition, creating a simple and open-source crowd estimation model is also a hidden requirement for social welfare. With the mentioned challenges, moving camera surveillance proves to be very successful in places where static cameras are not installed due to any power, technical or geographical issues. Moving camera surveillance proves to be useful for our security to keep an eye on the enemy in difficult or high-altitude places.

Rapid urban population growth is a serious matter of concern to us, which is being worked upon by the research community nowadays. According to the world demographic report, 55% of the world's population today lives in urban areas, which will increase by about 70% in 2030 [3], as a result, our future will be surrounded by an unstructured and unbalanced crowd [4]. Stampedes, intentional gun firing, mob lynching, unnatural accidents, and unstructured traffic can have significant consequences for such disorderly population growth [5]. Moreover, the frightening worldwide casualties from 1975 to 2019[1] inspired us to develop an efficient crowd counting mechanism in dynamic camera video surveillance environments. Modern CC approaches failed in accurate prediction of the crowd, resulting in lower accuracy; which also prompted us to work on advancing crowd counting via camera video surveillance.

Consecutive frame difference, context background subtraction, and optical flow are some traditional object detection techniques [6], [7], which often failed to handle some advanced pixel-level issues in motion parallax, moving background objects, blurring, and night vision. Optical flow is more robust among all but takes a longer time to produce real-time information than background subtraction [8]. Therefore, researchers are developing efficient and high-performing

* Corresponding author.
E-mail addresses: 87kumar.ankit@gmail.com (A. Tomar), amu.santosh@gmail.com (S. Kumar), pantbhasker@gmail.com (B. Pant).

[1] https://publications.iom.int/system/files/pdf/wmr_2020.pdf

Please cite this article in press as:
A. Tomar, S. Kumar, B. Pant. PeopleNet: A Novel People Counting Framework for Head-Mounted Moving Camera Videos, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.04.002

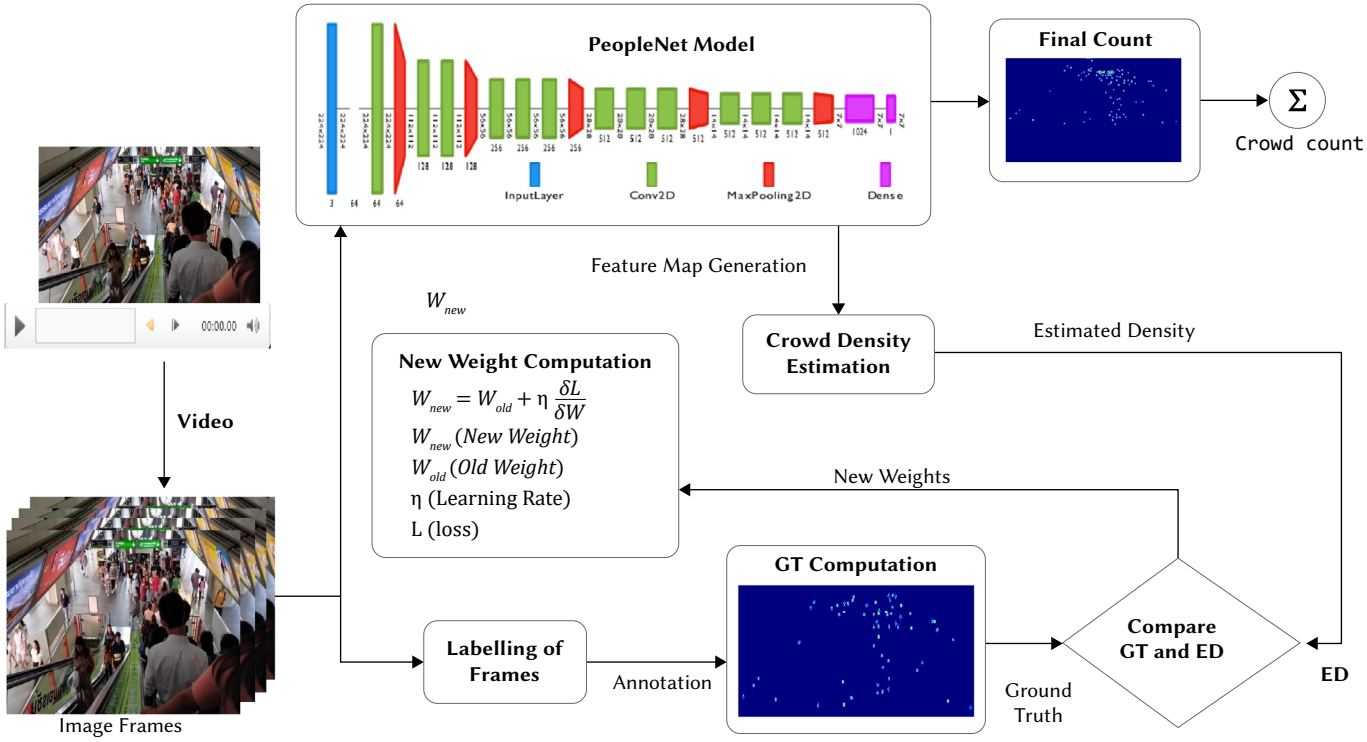

Fig. 1. Abstract working of People Counting framework.

methodologies [9], [10] for automatic crowd counting nowadays via density map (DM) generation due to its wide spreading and exploring the video surveillance domain.

This study addressed the deep learning-based PeopleNet framework to handle the stated challenges for crowd counting in adaptive tracking camera environments. This framework follows a transfer learning CNN mechanism to perform pedestrian detection, which is designed by disabling the upper convolution blocks and replacing the last layers with new fully connected, and dense layers in the standard VGG16. Some significant feature extraction techniques are applied before inputting the crowded frames for generating good quality density maps (shown in Fig. 1). The construction of a crowd dataset in a moving camera environment and developing a lightweight, vision-based [11] pedestrian estimation framework is a significant contribution to this work.

The rest of the paper is organized into four sections. Related work is presented in section II. The proposed system along with detailed experimental work is discussed in section III. The obtained crowd counting results and state of art discussion are carried out IVth section, and finally, the paper is concluded in section V.

II. Related Work

Based on the performance, object detection techniques are divided into motion detection and motion estimation categories. The current literature work illustrates motion detection methodologies in traditional and advanced research levels, which is further described in traditional and advanced object counting methods.

A. Traditional Object Counting

Early object detection work accomplished by watershed segmentation algorithm [12], often suffers from feature spatiality and complexity issues, further solved the spatial segmentation by deploying color quantization via edge-preserving techniques [13]. Markov Random Field (MRF) model got better segmentation results [14], [15]; but limited to stationary video sequences. Jordan et al., [16] proposed an object tracking and detection scheme for both fixed and moving camera videos; however, it failed to solve the spatial ambiguity issue. Traditional object detection schemes are easy to implement but not intensive to advance challenges such as luminous variation [17], dynamic appearance [18], abrupt motion [19], occlusion [20], [21], complex background [22], [23], shadow [24] and camera motions [25] etc. The researchers embed transfer learning networks in traditional object detection techniques nowadays to handle these issues.

B. Advanced Object Counting

Deep Neural networks (DNNs) are more popular since 2016 for foreground detection [26], background subtraction [27], background generation [28] and deep spatial feature extraction [29], [30]. Guo and Qi [31] first developed Restricted Boltzmann Machines (RBMs) for moving object detection using background subtraction mechanisms. A deep auto-encoder was used by Xu et al. [32] for object detection in moving camera images; on the other side, a context encoder was presented by Qu et al. [33] for background subtraction as a backbone. Droogenbroeck [34], Cinelli [35] and Bautista et al [36] used CNNs for background subtraction.

Recent studies used two-stage [37], structured [38] and cascaded [39] CNN etc. Whereas a (NeREM) Neural Response Mixture and Mixture of Gaussian (MOG) [18] framework are employed to learn deep crowd features. Lempitsky et al. [10], first converted the labeled images into density maps (DMs) with the sum of a fixed Gaussian Kernel. Y Zhang et al. [40], proposed a fixed standard deviation to generate density maps. A single-column CNN [41] uses ResNet50 as a backbone for feature extraction. Boominathan et al., [42] introduced a patch-based MCNN network, which consists of each column with a different kernel size of the same depth as each parallel column. This model is able to produce high quality of density maps in pure deep learning environment; however fails to compute the outcome in efficient time. CSRNet [43] further overcame the stated problem via using two parallel shallow and deep networks to maintain the original density map resolution.

Algorithm 1. Ground truth Generation

Input: A directory of .jpg files

T = 0, I_c;

Function `ReadImage`(*readpath*, $(T + i)$, I_c);

return readpath;

Function `ImageResize` (*l*, *b*);

l ← HeightOfImage;

b ← WidhtOfImage;

return resizedImage;

Function `WriteImage` ((*resizedImage*, [*readpath*, $(T + i)$, *filepattern*)], *filepattern*);

return writepath;

Function `SaveMatFile` (*writepath*);

return mat;

while $i = 1$, *to n* **do**

call ReadImage();

call ImageResize();

call WriteImage();

[x, y] ← getCoordinates;

imageInformation.location ← [x, y];

imageInformation.member ← [x, 1];

call SaveMatFile();

end

Output: A directory of .mat files

III. The Model Framework

High computation power requirement is one of the significant limitations of existing object detection mechanisms to tackle the dynamic background modeling implementation in real time. Lowering the computational overhead to get efficient crowd estimation results is a primary concern in the crowd counting techniques which had not performed efficiently for dynamic camera crowd videos. To tackle this issue of existing models, the PeopleNet model has proposed in this work. This model comprises the following components to estimate the pedestrian of the given video datasets.

1. Dataset Characteristics
2. Feature Generation
3. Feature extraction
4. The PeopleNet Architecture
5. Network Training
6. Crowd Counting
7. Performance Metrics

A. Dataset Characteristics

Promising work has been done in crowd counting form free and surveillance crowd samples [44]. Object detection problems can be solved in open CV, keeping static cameras into consideration; however not been practiced yet for dynamic objects. The unavailability of tracking camera video datasets is a significant need of the current research trend. So, we are obliged to construct an 1101 RGB image database prepared from the videos shot by a static or moving device. These videos have the different moving effects of the observing device and pedestrians. The dataset samples were shot in different places such as a mall, street, company corridor, restaurant, highway, escalator, roadside, tunnel, etc. in different timings, luminous appearances, shadows cast, in near-zero visibility that makes the dataset more realistic, practical, and overwhelming than existing ones. The crowd samples of videos of constructed datasets made public[2] for further research, of which the analogical description is shown in Table I.

TABLE I. Analogical Description of the Constructed Dataset

Dataset Attributes	Values
Resolution	1080x1920
Frame Rate (per second)	0.75
Total Samples	1101
Test/Train Samples	220:880
Crowd Variation	0-125
Total Pedestrians	33567
Color/Format	RGB/JPG
Place	Multiple Locations
Property	Walking, Eating
Average Crowd Size	31
Shadow/Reflection/Loitering	Yes

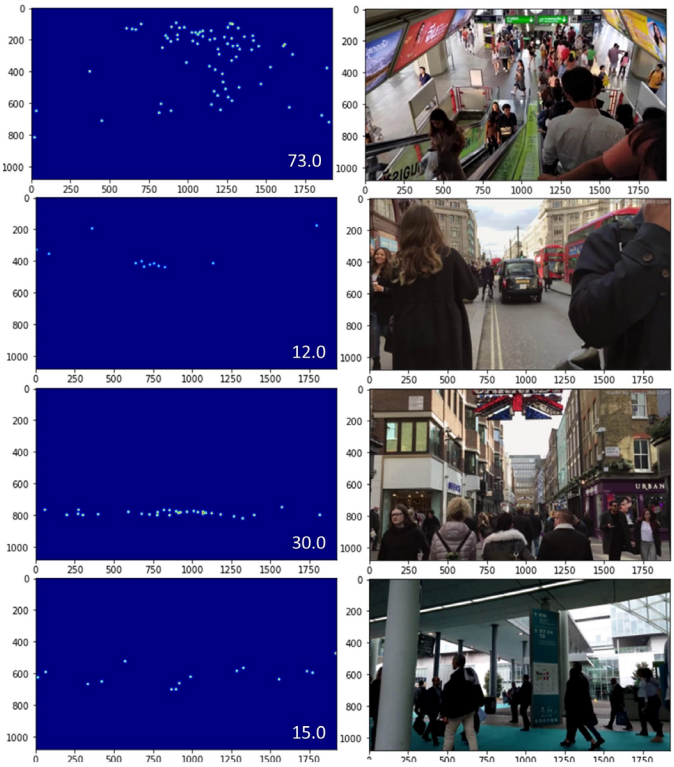

Fig. 2. Crowd samples with their respective density maps.

B. Feature Generation

The feature generation is an indivisible part of data preprocessing before input the information. The crowd features are generated via ground truth and density map generation, its whole process has discussed through pictorial and algorithmic way.

Ground Truth and Density Map Generation: Ground truth generation from the crowded images is based on current research trends [45], [46]. This work formulates the people estimation via density map via density regression function. The image frames and respective DMs (obtained from crowd video clips) are the training input of PeopleNet model. A DM of object detection is obtained from x, and y coordinates

[2] https://www.kaggle.com/ankit87/moving-camera-dataset

of the people's head location also called GT labels. The original image (bottom right portion), annotated samples (Left portion), and DM (Upper Right portion) are shown in Fig. 3. Consider the $I_i = (I_1, I_2, , I_{TS_n})$ are the image frames obtained (training samples) from crowded videos. The ground truth label $n^{G_T,i} = (n_1, n_2, ..., n_{X_i})$ for center point X represents each people's head presented in crowded samples; which is obtained via density map $D^{G_T,i}$, described through (1).

$$\forall p \varepsilon I_i, D^{G_T,i}(p) = \sum_{p \varepsilon n_i^{G_T}} G\, D^{G_T}(p; \mu = n^{G_T,i}, \sigma^2) \tag{1}$$

The GD^{G_T} is Gaussian distribution (σ) for 'p' pixels. The total crowd count in sample Ii is obtained by summing the density values for all pixels described in (2) and samples are shown in Fig. 2.

$$G_T = \sum_{p \varepsilon I_i} D^{G_T,i}(p) \tag{2}$$

Fig. 3. Obtained Image samples from Video (Right Lower), Annotation Creation (Left), and, generated Density Map(Right Upper) from the image sample.

C. Feature Extraction

To tackle the poor crowd counting performance due to high computation power demand, this work incorporated the focus of expansion (FOE) concept in feature preprocessing. The FOE plays a significant role in accurate flow estimation for CV applications such as range & obstacle estimation. The field effects of FOE signify the transformation and rotation motion caused by the dynamic camera. Efficient crowd feature (segment, edge-based, and texture) selection incorporating the feature of the expansion concept via diverging optical flow vectors to estimate the motion fields depicted in Fig. 4.

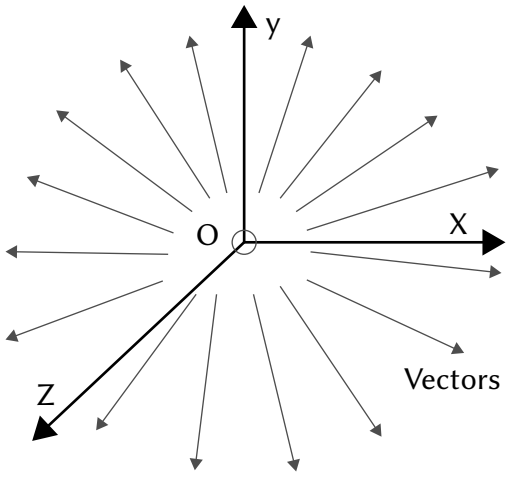

Fig. 4. FOE Diverging Optical flow vectors in temporal crowd.

Consider the $\vec{V} = (R_x, R_y, R_z)^T$ camera motion towards $P = (x, y, z)^T$ fixed points, where FOE is computed (x_{FOE}, y_{FOE}) against the pixel corresponding to the P crowd image plane. At a particular point of an image, FOE is obtained by the intersection of the image plane and camera motion velocity, while the camera is in relocatable motion [47].

$$V_x = \frac{R_z x - R_x f}{Z}, V_y = \frac{R_z y - R_y f}{Z} \tag{3}$$

A 3D coordinate system defines X, Y, and Z planes with the optical axis of the camera, which is parallel to the Z-axis and, X, Y-axis is parallel to the image plane for a particular location L(X, Y, Z) in a 2D plane at projection P(x, y, z) for the 3D plane [48]. The velocity V is derived under 3D space defined in (3), where V_x, V_y are velocity plane vectors. R_x, R_y, R_z are relocatable 3D components for focal distance f. Defining FOE in (4)

$$x_0 = \frac{fR_x}{R_z}, y_0 = \frac{fR_y}{R_z} \tag{4}$$

$$V_x = (x - x_0)\frac{R_z}{Z}, V_y = (y - y_0)\frac{R_z}{Z} \tag{5}$$

The (3) becomes in a linear system with (x_0, y_0) focus of expansion in (5) further used in (6).

$$\begin{bmatrix} V_{Y_1} & V_{Y_n} \\ V_{X_1} & V_{X_n} \end{bmatrix} * [x_0, y_0] = \begin{bmatrix} V_{Y_1}V_{X_1} & V_{Y_n}V_{X_n} \\ V_{Y_k}V_{X_k} & V_{Y_{nk}}V_{X_{nk}} \end{bmatrix} \tag{6}$$

The FOE detection is based on the above properties including flow and matched filter with size $(2w+1) \times (2w+1)$, in Fig 4, having each pixel shows the angle between the origin and grid point (7).

$$f(x, y) = archtan\left(\frac{x}{y}\right) - w \le x \le w, -w \le x \le w \tag{7}$$

For the given images $I_1(x,y)$ and $I_2(x,y)$, $\Delta t \to 0$ time apart, assume FOE based optical flow can be obtained corresponding to the flow of x, y-axis. Furthermore, the optical flow is tuned with segmented, edge-based, and texture features.

- **Segmented Features**: The segmented features capture foreground entities (blob, shape and size) at reference pixels for density map $D^{G_T,i}$, of mathematical expression is described in (8).

$$S = \sum_{n=0}^{p} S_n, where S_n = \sum_{(x,y) \in P_n} \sqrt{D^{G_T,i}(x,y)} \tag{8}$$

- **Edge-Oriented Features**: Consist Minkowski dimensions to estimate strong crowd counting ability via (9).

$$e = \sum_{n=0}^{p} e_n, where \;\; e_n = \sum_{(x,y) \in P_n} \sqrt{e(x,y)} \tag{9}$$

- **Local-Texture Features**: These features are employed for density classification across the crowded regions r, depicted in (10).

$$g(r) = \sum_{(x,y) \in r_n} 1, for Q_z(x,y) = Q_i(x', y') \tag{10}$$

D. The PeopleNet Architecture

In deep learning, CNN is enough capable of automatic feature extraction and prediction process, which evolved into transfer learning. The PeopleNet model is capable of crowd estimation using five convolution groups of 21 layers, initially, image samples were provided in batches with three RGB channels, as shown in Table [tabsecond]. The baseline network of this work is obtained from VGG16 architecture by stacking max pooling, convolution, dense, and fully connected (FC) layers. Two significant changes have been made to the standard VGG16 network to make it fit for people counting

purposes, first, we have disabled its top seven layers by freezing them (making their status 'false') and added our dense layers by replacing its FC layers. The input videos have an original resolution of 1080×1920, which further took 224×224 after preprocessing and normalizing. The input stream passed through Conv2D groups described by the following changes:

1. For the 1st Conv2D group, double convolution layers with 64 filters of [3×3] with a stride of (1, 1) have been applied followed by a 2D pooling layer of size [2×2] with stride (2, 2). However, we have frozen the first convolution group by setting their status '*False*'.
2. In 2nd Conv2D group, a double layer with 128 filters of [3×3] and stride (1, 1) have convolved over the output of the previous layers. We obtained a 2D pooling of 128 kernels after applying a stride of size (2,2).
3. The 3rd Conv2D group is composed of three layers having 256 filters of [3×3] dimensions with the stride of (1, 1). A stride (2, 2) and a max-pool [2×2] layer have been used in the sub-sampling technique.
4. The 4th Conv2D group has three consecutive convolution layers [3×3] with 512 filters in each and a stride (1,1) with pooling layers of size [2×2].
5. Likewise, the last and 5th group has convolution layers of 512 filters that have been convolved three times with the stride of (1, 1).
6. Finally, two fully connected layers of size 1024 and 1 are deployed at the end.

The proposed model works with n image frames I of M_i dimension matrix. A kernel K matrix is convolved through each image to create feature maps through equations (11) and (12).

$$s(i,j) = \sum_m \sum_n I(m,n)K(i-m)(j-n) \tag{11}$$

where

$$I[m,n] = \sum_m^s \sum_n^t (X[m+s][n+t]).C[s][t] \tag{12}$$

Where y is the output image, I image frame, C convolution mask, and t tokens. The value of K is taken 3 to carry out this experiment for $W \times H$ image dimensions for the p pooling matrix. The Euclidean loss is replaced by average pooling in (13) to estimate the N_i^{GT} as ground truth for spatial units U.

$$N_i^{G_T} = \frac{1}{U} \sum_j \hat{y}_i(x_j) \tag{13}$$

A customized loss function is required to train the effective DL models, which is derived from the difference between actual and estimated count (EC), depicted in (14).

$$W_{new} = W_{old} - \eta \frac{dL}{dW} \tag{14}$$

Where (x_i, y_i) are spatial coordinates for density map. W_{new} and W_{old} are the updated and older neuron weights gained at each forward and back prorogation with the help of learning rate η. The actual and estimated people counts are used to compute the pixel-level Euclidean distance loss function $L_D(\theta)$, which is defined in (15).

$$L_D(\theta) = \frac{1}{N_{ts}} \sum_{i=1}^{N_{ts}} \left(\frac{H_C)(I_i;\theta) - G_T}{G_T + 1}\right)^2 \tag{15}$$

A set of learn-able metrics (θ) derived from total sample parameters are N_{ts}. The actual head-count is denoted by G_T, and the estimated head-count is denoted by $H_c(I_i;\theta)$. The PeopleNet is trained from the scratch for random network parameters and resolves poor performance issues for a sparse crowd using $L_D(\theta)$, which helps to meet the real-time computation. The loss in (16) is given by

$$\nabla_\theta L_D = \{0; if N_i^{G_T}(I_i) - N_i^{G_T}(I_{i-1}) + \varepsilon \le 0 \nabla_\theta N_i^{G_T}(I_i) - N_i^{G_T}(I_{i-1}) \tag{16}$$

A popular mean of square error (MSE) loss is used to train the model via computing it between estimated and real pedestrian values. The experiment conduction used *Adam* as an optimizer function since it is best suitable for non-moving objects for noisy/sparse gradients, also provides a regret bound on convergence rate comparable to the convex optimizer eloberated in below equation (17).

$$\Theta_{t+1} = \Theta_t - \frac{\eta}{\sqrt{\tilde{v}_t + \epsilon}} \hat{m}_t \tag{17}$$

Here $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$ and $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$, are initialized vectors and biased towards 0, where β_1, β_2 are close to 1, whose default values are taken 0.9 and 0.999 respectively. The density map learning paradigm D(p) uses a set of X and Y parameters via mapping function F using θ as a parameterization function to predict the labels in (18).

$$D(p) = \int (F(X,\theta) - N^2) p(X,N) p(X,Y) \tag{18}$$

After each layer, the output is defined in (19); where O_i is output f is filtering, P is padding unit, and S is stride.

$$O = \frac{O_i - f + 2P}{S} \tag{19}$$

The proposed methodology is unique and uses shortcuts to all preceding blocks from each convolutional group. As a result, a combination of local and global features is fed to the first and last layers. We included padding, stride averaging, and regularization techniques for feature map matching.

E. Network Training

CNN-based people counting frameworks have many trainable parameters or complex architectures, which makes model training more difficult and the feeding process time-consuming. In this study, the PeopleNet model is performing better on fewer trainable parameters in less computation. For the given training set of 880 images and their respective DMs, two counter variables F and G train the convolution model jointly with the help of $L_D(\theta)$. Algorithm 2 summarizes the training procedure for the proposed model. Let (I_i, D_i) be the pair of i^th image and density maps, and (F_{l_i}) (G_{D_i}) are original and refined density maps respectively, the loss (L) can be expressed through (20).

$$L(\theta)_D = \sum_{i=1}^{N_{ts}} |F(I_i) - G(D_i)|^2 + \alpha |G(D_i) - D_i|^2 \tag{20}$$

$|G_{D_i} - D_i|$ helps in training and predicting the refined density maps during every back and forward prorogation, and the loss tends to converge to the minimum level for smooth training. The training accuracy and validation loss statistics for 300 epochs of PeopleNet are shown in Fig. 5.

F. Crowd Counting

The position of each person's head is labeled with white cross symbols (H_x, H_y) as a delta function (Fig. 3); which is used to compute the labeled images into a DM by convolving operation through G_{α_i}. The inverse KNN distance method obtains labeled values from G_T by computing KNN distance from $p(X_x, X_y)$ to $p(X_x, X_y)$ pixel values, for H people heads (depicted in algorithm 2). The data samples and respective annotated DMs are parallel inputs provided to the PeopleNet model, which can generate final maps after 8 hours of intensive training.

TABLE II. Layered Architecture of PeopleNet Framework.

Layer Type	Output Shape	Parameters	Trainable Status
Conv2D	64 × 64 × 224 × 224	1792	False
BatchNorm2D		128	False
Relu			
Conv2D		36928	False
BatchNorm2D		128	False
Relu			
MaxPoll2D			
Conv2D	64 × 128 × 112 × 11	73856	True
BatchNorm2D		256	True
Relu			
Conv2D		147584	True
BatchNorm2D		256	True
Relu			
MaxPoll2D			
Conv2D	64 × 256 × 56 × 56	295168	True
BatchNorm2D		512	True
Relu			
Conv2D		590080	True
BatchNorm2D		512	True
Relu			
Conv2D		590080	True
BatchNorm2D		512	True
Relu			
Conv2D	64 × 512 × 28	1110860	True
BatchNorm2D		1024	True
Relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
MaxPooling2D			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
MaxPooling2D			
AdaptiveAveragePooling2D			
Flatten			
BatchNorm2D		2048	True
DropOut			
Linear	64 × 512	524288	True
Relu			
BatchNorm1D		1024	True
DropOut			
Linear	64 × 1	512	True

Algorithm 2. PeopleNet Training Procedure

Input: A pair of Images and respective density maps $(I_i, D_i)_{i=1}^{N}$

Initialize two F and G counters;

for *Epoch* ← 1..... N_E **do**

  for *Epoch* ← 1..., .. N **do**

    Estimate DM (F_{l_i}).

    Generate G_T (G_{D_i}).

    Update F counter using loss L in (20)

    update N_G for every epoch through counter G

    if mode(*Epoch*, N_G) == 0 **then**

      update the parameter using $L_D(\theta)$, (15)

    end

    else

    end

  end

end

Output: Updates values of counters F and G.

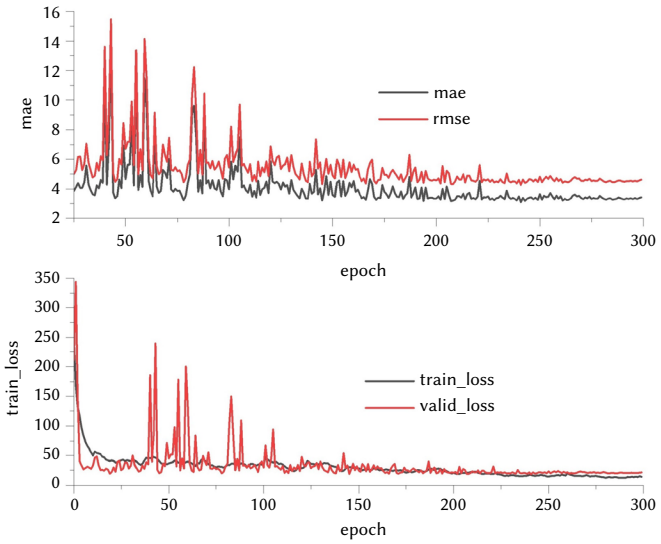

Fig. 5. Training accuracy and loss statistics of PeopleNet.

In the experimental procedures, it has been observed that the obtained result is more precise upon disabling the first convolution layers than the other layers for the stated loss function (Fig. 6). In this figure, the first two rows have the original image and original DMs, the next rows show the output generated DMs (printed with sequence, PSNR, G_T, and E_C).

G. Performance Metrics

After widely investigating the universally accepted articles, we have divided the performance evaluation into the image and pixel-level categories [24]. The quality of generated DMs is used to evaluate the pixel-level performance, whereas popular regression model enumerate the image-level performance.

1. Image Level Error

Root Mean Squared Error (RMSE) and, Mean absolute error (MAE) are commonly used as image-level accuracy measurement metrics [49], which compute the overall deviation between the estimated and actual samples values. The mathematical expressions of RMSE and MAE can be seen in (21) and (22).

$$RMSE = \sqrt{\frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}(C_i^{G_T} - C_i^{E_C})^2} \tag{21}$$

$$MAE = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}(C_i^{G_T} - C_i^{E_C}) \tag{22}$$

RMSE is a more corrective measure to assess the insignificant sample deviation; however, MAE generally fails to secure the overall accuracy for huge variation data samples. Therefore, mean absolute percentage error (23) would be the better measurement choice.

$$MAPE = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}\left|\frac{C_i^{G_T} - C_i^{E_C}}{C_i^{G_T}}\right| \tag{23}$$

MAE, MAPE, and RMSE define robustness and global image-level accuracy. To evaluate local region accuracy, correlation coefficient r is another better alternative to measure the dependence between two variables (24), where $A = C_i^{G_T}$ and $B = C_i^{E_C}$.

$$r = \frac{N_{ts}(\sum(A)(B)) - (\sum(A))(\sum(B))}{\sqrt{[N_{ts}\sum(A)^2 - (\sum(A))^2] * [N_{ts}\sum(B)^2 - (\sum(B)^2]}} \tag{24}$$

To measure the covariance, association or statistical relationship of two continuous variables, the Pearson correlation coefficient (P_r) is another significant test statistics, which is the fundamental measure the strength of the linear relationship shown in (25). Sometimes it is also called as coefficient of determination.

$$P_r = \frac{\sum_{i=1}^{N_{ts}}(C_i^{G_T} - \overline{C_i^{G_T}})(C_i^{E_C} - \overline{C_i^{E_C}})}{(\sqrt{\sum_{i=1}^{N_{ts}}(C_i^{G_T} - \overline{C_i^{G_T}})^2}\sqrt{\sum_{i=1}^{n}(C_i^{E_C} - \overline{C_i^{E_C}})^2}} \tag{25}$$

Adjust R_A^2: R^2 often suffers from score improvement in increasing terms, even if the model improvement remains the same, which might create a misguidance for the researchers. Therefore, R_A^2 is used to improve in case of any real improvement via adjusting the increased estimators, in (26) k is independent variables for n observations in operation.

$$R_A^2 = 1 - (1 - R^2)\left[\frac{N-1}{n-1-k}\right] \tag{26}$$

The Normalized Root Mean Square Error (NRMSE) (27) use to facilitates the comparison between models with different scales to interpret as a fraction of the overall range that is typically resolved by the model.

$$nrmse = \frac{RMSE}{\overline{C_i^{E_C}}} \tag{27}$$

$\overline{C_i^{E_C}}$ is the average of observation value computed (21).

2. Pixel Level Error

PSNR (Peak signal-to-noise ratio) [50] is the most common pixel-level metric used to measure the error deviation between corresponding and original DM pixels. The resolution size of both the original and degraded image matrix must be the same while working with the 2D matrix.

$$PSNR = 10log_{10}(\frac{MAX_f}{\sqrt{MSE}}) \tag{28}$$

In (28), 10log defines the square of amplitudes in terms of noise and, MSE is defined in (29).

$$MSE = \frac{1}{M * N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}|O(i,j) - D(i,j)|^2 \tag{29}$$

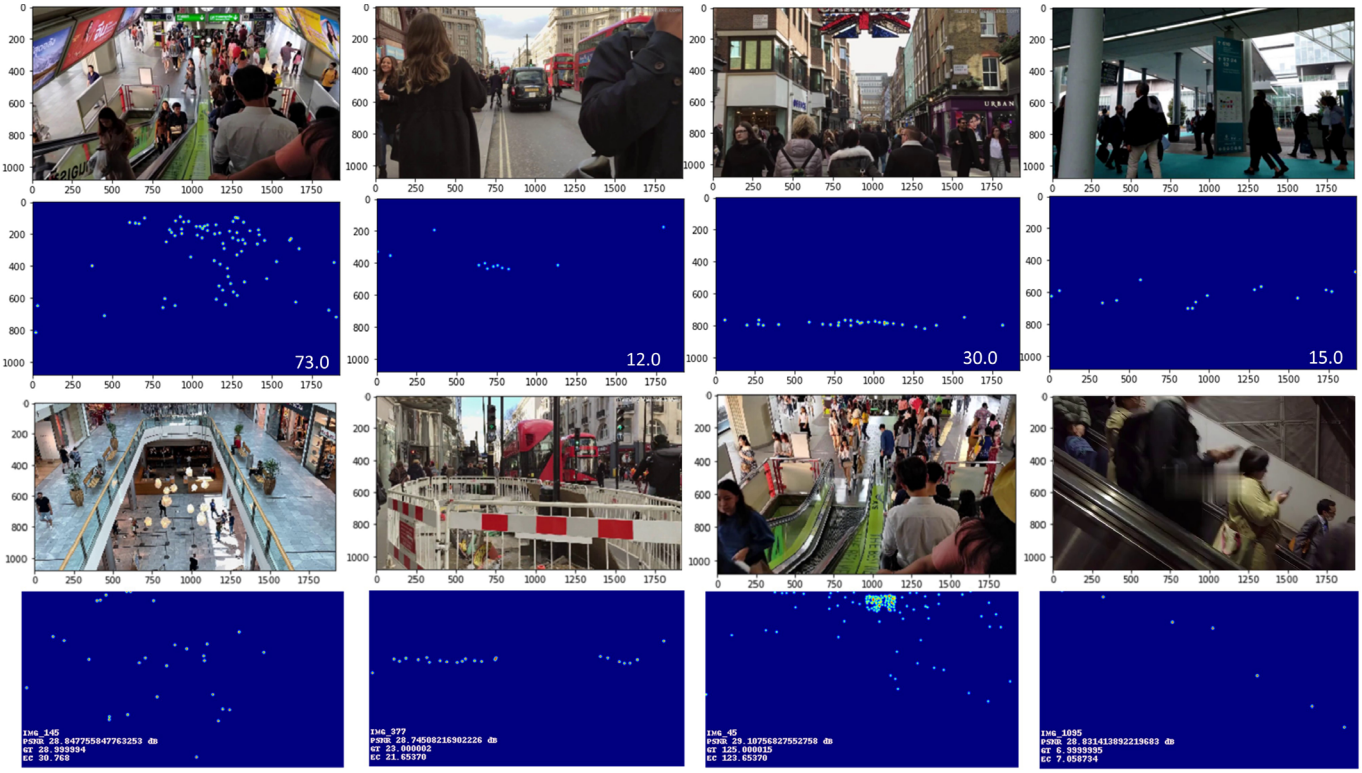

Fig. 6. Frames and their respective density maps are shown in first two rows, while last two rows represent testing output frames and generated density maps printed with some information viz image-sequence, PSNR value, GT and EC, etc.).

TABLE III. Performance Analysis of Proposed Model Different Category Samples

Category	Total Frames	Crowd Size	MAE	RMSE	Pr	r	nrmse	SSIM
Dense	165	51-125	15.514	19.595	0.890	0.944	0.124	0.31
Sparse	260	0-15	12.24	15.453	0.912	0.932	0.131	0.37
Average	676	16-50	9.376	12.72	0.907	0.933	0.156	0.35
Overall	1101	0-125	3.43	4.623	0.917	0.919	0.166	0.34

Where O and D are data matrices of original and degraded DMs. MAX_f denote maximum signal value for M rows denoted by i and N columns represented by j pixels. Structural Similarity Index: SSIM([50]) is used to measure the perceptual difference between two identical images this parameter cannot judge the better of two images.

IV. Result and Discussion

This section mainly focuses on the validation efficacy of PeopleNet architecture for a variety of crowd scene categories. The training and accuracy error deviation is expressed separately through Fig. 5 and Fig. 7. Moreover, the comparison of the cross-scene results is comprehended separately for existing state of art CC models and benchmark datasets.

A. PeopleNet Model Performance Analysis & Validation

The crowd estimation results are validated over the test data samples, whose predicted results are computed and shown graphically. A thorough, deep, and strategic error deviation of this experiment has been done in three main parts as follows:

1. Scenario Error Estimation

Table III presents the results obtained from the PeopleNet framework on the constructed dataset. The data samples have been categorized into dense, sparse, and average scenarios. The framework is tested separately by modeling in each category to better analyze the people counting error. The dense category data subset contains 165 frames having a people range from 51 to 125, over which the PeopleNet model secures the MAE, RMSE, Pr , r, nrmse and SSIM as 15.514, 19.595, 0.890, 0.944, 0.124 and, 0.31 respectively. The sparse crowd category kept a total of 260 samples (having people ranging between 0 to 15) and the model secured 12.24, 15.45, 0.912, 0.932, 0.131 and 0.37 as MAE, RMSE, *Pr*, *r*, nrmse and SSIM respectively. Moreover, there are 676 average category samples of 16-50 crowd size and secured 9.376, 12.72 as MAE and RMSE.

Satisfactory performance is observed for sparse category samples; insufficient training samples are the main cause of lower accuracy. Instead of training over enough samples, the model achieved satisfactory results for average crowd samples by securing average MAE and RMSE. This accuracy is low as compared to the overall category dataset. The proposed model performance is superior in the overall category dataset sample with 1101 frames of people ranging from 0-125 and achieved MAE 3.43 and RMSE 4.623 respectively with SSIM 0.34. A slight difference is observed in model performance from various crowd categories, which indicates that the proposed model neither performs poorly for dense nor outperforms for sparse samples. The main aim to present the Table III is to test the PeopleNet's robustness for low and high-crowded samples.

2. Training Error Estimation

Existing CC research focused on the accuracy parameters only, but the proof of how well the model training performed is still missing in the literature. Here we exposed the training performance through pictorial and factual representation. During training, the difference

between MAE and RMSE is higher in the first 50 epochs; however, it is less afterward but fluctuates with a very high margin. The fluctuation lowered between 100 to 200 epochs. After 200 epochs, a negligible fluctuation is observed as both MAE and RMSE are parallel to one another from 150 epochs till the end. These statistics ensure a smooth training process without over, or under-fitting.

3. Crowd Error Estimation

The performance of any AI model mainly depends on its architecture, hyperparameter setting, quality, and quantity of training samples. In this experiment, the obtained results vary from test-to-test samples as they may belong to a different video clip. The accuracy deviation between estimated and real people obtained by the proposed model over 220 different image frames could be seen in Table IV.

TABLE IV. Summary of Accuracy Deviation Obtained by the Proposed Model Over 220 Image Frames

Frame Number	Frame Image	Ground Truth	Estimated Count	Deviation Accuracy
1^{th}	f_1	29	26.383	2.617
2^{nd}	f_2	29	31.4189	-2.4189
——	——	——	——	——
10^{th}	f_{10}	39	49.2062	-10.2062
11^{th}	f_{11}	06	7.3051	-1.3051
——	——	——	——	——
50^{th}	f_{50}	14	12.0589	1.9411
51^{th}	f_{51}	27	27.4985	-0.4985
——	——	——	——	——
100^{th}	f_{100}	69	64.2111	4.7889
101^{th}	f_{101}	23	27.5823	-4.5823
——	——	——	——	——
150^{th}	f_{150}	101	86.4342	14.5658
151^{th}	f_{151}	04	6.0725	-2.0725
——	——	——	——	——
200^{th}	f_{200}	31	28.1497	2.8503
201^{th}	f_{201}	23	22.8873	0.1127
——	——	——	——	——
119^{th}	f_{119}	03	05.1674	-2.1674
220^{th}	f_{220}	41	48.7397	-7.7397

Furthermore, the same results are also depicted through Fig. 7, where a yellow slider area is mapped to render the zoomed view. For each edge, the quantified results have been associated, the absolute value represents C_i^{GT} and the fractional values represent C_i^{EC}. The underestimation CC effect can be observed for more than 100 people. Insufficient training samples for dense crowd samples may be the leading cause; however, the nearness of obtained results and parallelism of lines is evidence of the model's outstanding results.

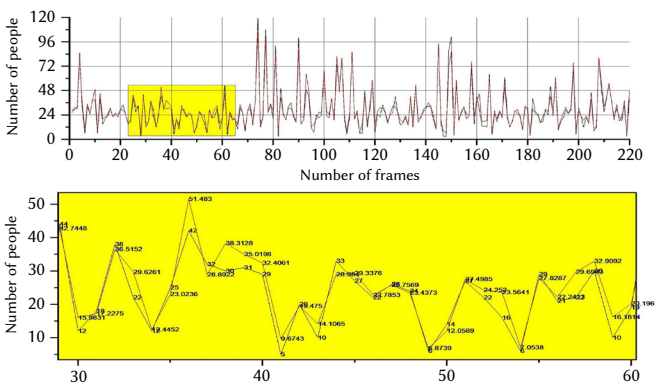

Fig. 7. Accuracy deviation illustration between estimated and actual crowd count.

B. Comparative Analysis

This work focuses on the result variation, model behavior, and data variance on moving camera datasets after cross-testing on one another crowd videos. The robustness and correctness of both entities are contentious, which could be observed in the following cases:

1. The PeopleNet generates promising results at pixel and image level over moving crowd scenes, which are are complementary to each other. Whether the constructed dataset fulfills the current CC research expectations? It is a question.
2. Table III, V and VI validates the correctness of data samples by exploring the model's capability in different crowd scenarios. Whether this novel framework is acceptable for public and standard crowd datasets? It is also a question.

We have answered the above questions in detailed along-with significant case studies 1 and 2.

1. Case Study 1

Testing the crowd counting ability of a novel model on a newly constructed dataset has always been challenging. Comparing the efficiency of some extant object counting models with the proposed methodology is an integral part of any comparative analysis research has been illustrated in Table V, which overviews the comparative performance analysis of existing CC models and the proposed model. As per our best knowledge, due scarcity of tracking camera surveillance crowd datasets, we are forced to compare the Constructed dataset for validation. Some of the universally agreed performance evaluation metrics have been computed for popular deep learning models such as ResNet50 [41], CSRNet [43], DENet [51] and, People-Flow [52] and presented in tabular form. These models are originally developed to perform the crowd counting in free or surveillance view crowd datasets captured by a static camera. Therefore, the performance of these CC models will deteriorate with the proposed database. A PSNR and SSIM values of 26 and 0.43 respectively ensure the high quality of density maps results in accurate people estimation. However, the correlation (C_r) and Pearson coefficient (P_r) variation were observed significantly across all samples. These variations have been received with different samples due to the architectural complexity of various models.

A dilated-CNN structure with 2,160,000 trainable parameters (290.024 MB) in CSRNet [43], is specially designed for density estimation on highly congested crowd datasets such as Shanghai-Tech [40], *WorldExpo* [41], *UCF_CC_50* [56] and UCSD [57] datasets. Its lightweight functionality along with the front (basic CNN) and back-end network is a significant advantage over existing models to generate high-quality density estimation with less hardware computational training effort. A decremented variation of 5-18% and 12-21% for MAE and RMSE has registered over four different random sub-samples, and a decremented variation of 40-75% has registered for the pixel level accuracy of generated DMs. A PSNR and SSIM values of 26 and 0.43 ensure the high quality of density maps responsible for accurate people estimation.

The ResNet-50 [41] model was tested over four random sub-data samples and secured a decrease of 55-80%, and a decrease of 30-40% in MAE and RMSE respectively. However, an increase of 40-50% has been observed for image level (PSNR) accuracy. This ensures the residual learning and strong feature extraction ability of the standard VGG19 which has been used as a backbone across 16,263,489 trainable parameters of ResNet50. The significant disadvantage of this network is to produce the low quality of density maps with poor regularization techniques, and as results secure lower PSNR and SSIM values. The percentage variation for image and pixel level accuracy is caused by training over lower training samples having mixing data (sparse and dense crowd). However, the lightweight training feature extraction ability is one of the significant advantages of this model, requiring

TABLE V. Summary of Comparing Performance Metrics of Different Frameworks With Overall and Segmented Portions (Viz Subset1, Subset2, Etc.) of Self-Constructed Datasets. The Proposed Framework PeopleNet Has Marked With . CSRNet [43] & DENet [51] Frameworks Are Tested Over Two Data Subsets Due to Having Heavily Loaded Front End and Back-End Integrated Networks. The Down Arrow (↓) Means the Lower the Metrics Higher the Accuracy, and the up Arrow (↑) Means the Higher the Metrics Higher the Accuracy. The Term 'NA' Denotes Not Available

Category		SampleCount	PerformanceMetrics								
			ImageLevel						PixelLevel		
Models	**Test Subsets**	**(20:80)/100**	**MAE(↓)**	**RMSE(↓)**	**MAPE(%)(↓)**	**PearsonCoefficient (Pr)(↑)**	**CorrelationCoefficient (r)(↑)**	**R^2_A(↑)**	**nrmse(t)**	**PSNR(dB)(↑)**	**SSIM(↑)**
CSRNet[43]	**Subset1**	110:441/551	17.89	22.43	NA	NA	NA	NA	0.133	13.21	NA
	Subset2	110:441/551	15.51	20.89	NA	NA	NA	NA	0.126	14.76	NA
	Overall	**220:881/1101**	**10.03**	**16.09**	**NA**	**0.922**	**0.931**	**0.941**	**0.117**	**25.99**	**0.43**
ResNet50[41]	**Subset1**	55:220/275	3.97	6.70	18.47	0.953	0.976	0.969	0.142	11.45	NA
	Subset2	55:220/275	3.64	4.89	21.70	0.886	0.942	0.910	0.199	9.67	NA
	Subset3	55:220/275	3.38	5.08	20.21	0.865	0.915	0.892	0.191	5.22	NA
	Subset4	55:220/275	3.38	4.28	21.64	0.856	0.925	0.877	0.181	9.22	NA
	Overall	**220:881/1101**	**4.14**	**5.49**	**17.53**	**0.888**	**0.934**	**0.903**	**0.176**	**19.87**	**0.30**
DENet [51]	**Subset1**	110:441/551	6.60	11.27	21.02	0.887	0.901	0.899	0.155	NA	NA
	Subset2	110:440/551	5.97	11.74	19.83	0.891	0.926	0.917	0.143	NA	NA
	Overall	**220:881/1101**	**4.58**	**6.12**	**18.78**	**0.788**	**0.947**	**0.805**	**0.152**	**NA**	**NA**
FlowNet[52]	**Overall**	**220:881/1101**	**3.23**	**5.33**	**17.25**	**0.81**	**0.905**	**0.822**	**0.152**	**NA**	**0.26**
***PeopleNet* ***	**Subset1**	55:220/275	3.76	5.23	25.10	0.974	0.987	0.988	0.124	39.61	NA
	Subset2	55:220/275	2.95	4.20	15.90	0.902	0.950	0.92	0.143	34.23	NA
	Subset3	55:220/275	3.25	3.45	17.90	0.922	0.932	0.941	0.157	31.21	NA
	Subset4	55:220/275	2.76	3.67	19.03	0.853	0.924	0.862	0.164	47.01	NA
	Overall	**220:881/1101**	**3.306**	**4.38**	**19.75**	**0.977**	**0.954**	**0.923**	**0.160**	**24.12**	**0.52**

TABLE VI. Results Validation on Various Datasets, Highlighted Black Text Represents the Results Obtained on Random Sub-Samples, Highlighted Blue Text Represents the Results Obtained by the PeopleNet Framework

Model	Dataset/ TestSubset		SampleSize (Test:Train)/Total	Performance	
				MAE	RMSE
*PeopleNet**	Mall[49]	Subset-1	100:400/500	1.648	2.116
		Subset-2	100:400/500	1.799	2.193
		Subset-3	100:400/500	1.407	1.708
		Subset-4	100:400/500	1.589	2.094
		Overall	**400:1600/2000**	**1.247**	**1.33**
SAAN[53]			**400:1600/2000**	**1.28**	**1.68**
*PeopleNet**	Beijing-BRT[54]	Subset-1	128:512/640	11.65	16.179
		Subset-2	128:512/640	8.933	14.325
		Overall	**256:1024/1280**	**3.172**	**4.1634**
DRResNet[54]			**256:1024/1280**	**1.39**	**2.00**
*PeopleNet**	Shanghai-Tech-B[40]	Subset-1	71:287/358	20.916	30.015
		Subset-2	71:287/358	25.583	42.108
		Overall	**143:573/716**	**9.807**	**14.265**
SPANet[55]			**143:573/716**	**6.50**	**9.90**

fewer neuron weights (after training 187 MB) to perform pedestrian detection for dense crowd images.

The People-Flow model [52] was used to perform pedestrian estimation via using standard VGG16. It has a front end and its layers as the back end to ensure high-quality density maps. This integrated architecture secures 3.2, and 5.3 as MAE and RMSE; which is good enough but 0.152 as nrmse ensures the model's robustness towards object head courting even in high crowd density samples while securing 0.26 as the SSIM value. The DENet [51] is an integrated detection (DENet) and estimation network (ENet) that performs CC tasks by using an encoder-decoder network as a dual-end network that trains separately over Mask R-CNN (9,64,983 trainable parameters). In comparison to 1101 samples for DENet, an increase of 30-44% in MAE and 85-95% in RMSE is obtained over two random sub-datasets; this shows vibrant performance for data variation. Due to the encoder-decoder architecture of DENet, it disables pixel annotations, and computing the PSNR coefficient is nearly impossible. This universal network achieved MAE and RMSE values of 4.23 and 5.67, respectively, but failed to generate high-quality DMs due to low crowd density.

As we can see in Table V, MAE, RMSE, and MAPE lower values support higher regression accuracy between actual and observed values. The MAPE has calculated by dividing the difference by the actual value, in which if the actual value is close to 0 then the error will be very high, so MAPE is to be used only when the actual value is far from 0. That is the main reason for securing the MAPE of the PeopleNet model is higher than FlowNet for complete data samples. The Pearson and correlation coefficients determine the relationship strength, the higher the value the stronger the relationship. The CSRNet, ResNet50, DeNet, and FlowNet architecture showed a moderate degree of correlation as securing the average Pearson and correlation coefficients over their subset and overall testing samples for crowd counting purposes. Always having a higher correlation never means a strong relationship as it is a bi-variate relationship that somehow depends on the network architecture also, which has observed in the case of subsets testing of CSRNet model rather than PeopleNet. The CSRNet model adds a high number of useful variables as compared to existing models which results in high adjusted R^2. However, the proposed PeopleNet model secures R^2 a total of 0.977 by clearing the misconception of lower regression accuracy for low to adjust R_A^2 instead of having fewer neurons as CSRNet. The nrmse indicator is not always reliable for finding the best networks for small training samples, which is indeed shown in the case of ResNet50. The proposed model traced higher 'nrmse' among the models utilizing front and back-end training architecture. On another hand, the pixel-level performance of the PeopleNet model is incomparable as compared to existing state of art crowd counting models. The better overall object counting accuracy of the proposed model ensures its scalability and robustness even in videos shot in extreme conditions.

2. Case Study 2

Table VI shows the acquired findings over the benchmark and public datasets to compare overall performance without bias. Various statistical reports and results were presented to differentiate the accuracy variance for different data subsets. The RMSE for each random subset having 500 samples in the Mall dataset fluctuates between 28 and 65% when compared to the entire data samples; nevertheless, SAAN [53] achieved a nearly 26% increase.

The sparse, dense sub-sampling disparity has shrunk marginally in MAE, but it is essentially non-existent in SAAN [53] and PeopleNet. Because the Beijing-BRT [54] includes a total of 1280 samples, we computed PeopleNet results for two random and equal subsets due to the smaller data samples. We can see the nuanced variance in MAE but not in RMSE in each case; however, the proposed model has demonstrated superiority over the DRResNet [54] model by obtaining near 50% reductions in MAE and RMSE. There are 716 samples in the ShanghaiTech-B [40]. As a result, we only assessed PeopleNet's performance on two random samples, subset1 and subset2. For the Beijing-BRT and ShanghaiTech-B samples, there was a 100-200% change in MAE and RMSE, with a smaller percentage loss for DRResNet [54] and SPANet [55] accuracy. The obtained results present some image-level performance on frequently used databases; however, the pixel-level accuracy comparison is not helpful for static devices captured in existing datasets. The obtained results by the PeopleNet model on the proposed dataset are closer to the obtained on existing datasets; which is solid evidence of the correctness, scalability, and robustness of the proposed model over constructed data samples.

V. Conclusion

Employing a novel PeopleNet framework, this study handled a difficult CV problem of autonomous crowd counting using a tracking dynamic imager. Using the feature of expansion residual mapping over the camera-induced motion for a self-generated head-mounted video dataset, this mechanism performs BEYOND and IN operations for visible spectrum. The technical aspect of this model is to provide fair people counting over moving cameras and moving people

without intentionally or unintentionally pointing out individuals. The behavioral aspect of this study includes human counting in dangerous enemy territory or isolated places where electricity and infrastructure are no longer available. The PeopleNet's experimental findings revealed that pedestrian recognition is done efficiently in the day or night environments to address occlusion. Detection of social distance practice violation via crowd density estimation could be another significant social aspect of this work.

Extend the PeopleNet model's functionality for Covid19 like virus protocols via crowd monitoring in public places will be the possible future scope of this work.

Acknowledgment

We are grateful to Graphic Era University to provide the computational resources execution of this research work. We also give special thanks to the research team and websites for providing the public Mall, SmartCity, and ShanghaiTech-B dataset repositories. We are also grateful to the anonymous reviewers for providing us with valuable comments to make this research work better.

References

[1] A. Ferligoj, V. Batagelj, "Direct multicriteria clustering algorithms," *Journal of classification*, vol. 9, no. 1, pp. 43– 61, 1992.

[2] H. Faris, I. Aljarah, S. Mirjalili, "Training feedforward neural networks using multi-verse optimizer for binary classification problems," *Applied Intelligence*, vol. 45, pp. 322–332, 2016.

[3] A. Korotayev, J. Zinkina, "Egypt's 2011 revolution: A demographic structural analysis," in *Handbook of revolutions in the 21st century: The new waves of revolutions, and the causes and effects of disruptive political change*, Springer, 2022, pp. 651–683.

[4] C. A. Martin, C. Marshall, P. Patel, C. Goss, D. R. Jenkins, C. Ellwood, L. Barton, A. Price, N. J. Brunskill, K. Khunti, *et al.*, "Association of demographic and occupational factors with sars-cov-2 vaccine uptake in a multi-ethnic uk healthcare workforce: a rapid real- world analysis," *MedRXiv*, pp. 2021–02, 2021.

[5] E. A. Felemban, F. U. Rehman, S. A. A. Biabani, A. Ahmad, A. Naseer, A. R. M. A. Majid, O. K. Hussain, A. M. Qamar, R. Falemban, F. Zanjir, "Digital revolution for hajj crowd management: a technology survey," *IEEE Access*, vol. 8, pp. 208583–208609, 2020.

[6] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, *et al.*, "A system for video surveillance and monitoring," *VSAM final report*, vol. 2000, no. 1-68, p. 1, 2000.

[7] M. Adimoolam, S. Mohan, G. Srivastava, *et al.*, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 112–120, 2022.

[8] A. Sobral, E.-h. Zahzah, "Matrix and tensor completion algorithms for background model initialization: A comparative evaluation," *Pattern Recognition Letters*, vol. 96, pp. 22–33, 2017.

[9] B. Xu, G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–8, IEEE.

[10] C. Arteta, V. Lempitsky, J. A. Noble, A. Zisserman, "Interactive object counting," in *Computer Vision– ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, 2014, pp. 504–518, Springer.

[11] C. Zhang, Z. Liu, C. Bi, S. Chang, "Dependent motion segmentation in moving camera videos: A survey," *IEEE Access*, vol. 6, pp. 55963–55975, 2018.

[12] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 539–546, 1998.

[13] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[14] A. Ghazvini, S. N. H. S. Abdullah, M. Ayob, "A recent trend in individual counting approach using deep network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 7–14, 2019.

[15] A. Ghosh, B. N. Subudhi, S. Ghosh, "Object detection from videos captured by moving camera by fuzzy edge incorporated markov random field and local histogram matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 8, pp. 1127–1135, 2012.

[16] P.-M. Jodoin, M. Mignotte, C. Rosenberger, "Segmentation framework based on label field fusion," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2535–2550, 2007.

[17] B. Lee, M. Hedley, "Background estimation for video surveillance," in *Image and Vision Computing*, 2002, pp. 315–320, CSIRO.

[18] C. Stauffer, W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition*, vol. 2, 1999, pp. 246–252, IEEE.

[19] O. Munteanu, T. Bouwmans, E. Zahzah, R. Vasiu, "The detection of moving objects in video by background subtraction using dempster-shafer theory," *Transactions on Electronics and Communications*, vol. 60, no. 1, pp. 1–9, 2015.

[20] C. Marghes, T. Bouwmans, R. Vasiu, "Background modeling and foreground detection via a reconstructive and discriminative subspace learning approach," in *International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV*, vol. 2012, 2012.

[21] R. C. Joshi, A. G. Singh, M. Joshi, S. Mathur, "A low-cost and computationally efficient approach for occlusion handling in video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 28–38, 2019.

[22] J. A. Ramirez-Quintana, M. I. Chacon-Murguia, "Self- adaptive som-cnn neural system for dynamic object detection in normal and complex scenarios," *Pattern Recognition*, vol. 48, no. 4, pp. 1137–1149, 2015.

[23] J. A. Ramírez-Quintana, M. I. Chacon-Murguía, "Self- organizing retinotopic maps applied to background modeling for dynamic object segmentation in video sequences," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8, IEEE.

[24] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[25] H.-x. Zhang, D. Xu, "Fusing color and gradient features for background model," in *2006 8th International Conference on Signal Processing*, vol. 2, 2006, IEEE.

[26] D. Zeng, M. Zhu, A. Kuijper, "Combining background subtraction algorithms with convolutional neural network," *Journal of Electronic Imaging*, vol. 28, no. 1, pp. 013011–013011, 2019.

[27] M. Babaee, D. T. Dinh, G. Rigoll, "A deep convolutional neural network for background subtraction," *arXiv preprint arXiv:1702.01731*, 2017.

[28] L. Xu, Y. Li, Y. Wang, E. Chen, "Temporally adaptive restricted boltzmann machine for background modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[29] M. J. Shafiee, P. Siva, P. Fieguth, A. Wong, "Real- time embedded motion detection via neural response mixture modeling," *Journal of Signal Processing Systems*, vol. 90, pp. 931–946, 2018.

[30] M. J. Shafiee, P. Siva, P. Fieguth, A. Wong, "Embedded motion detection via neural response mixture background modeling," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 837–844, IEEE.

[31] R. Guo, H. Qi, "Partially-sparse restricted boltzmann machine for background modeling and subtraction," in *2013 12th International Conference on Machine Learning and Applications*, vol. 1, 2013, pp. 209–214, IEEE.

[32] P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang, J. Ding, "Dynamic background learning through deep auto- encoder networks," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 107– 116.

[33] Z. Qu, S. Yu, M. Fu, "Motion background modeling based on context-encoder," in *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, 2016, pp. 1–5, IEEE.

[34] M. Braham, M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *2016 international conference on systems, signals and image processing (IWSSIP)*, 2016, pp. 1–4, IEEE.

[35] L. P. Cinelli, "Anomaly detection in surveillance videos using deep

residual networks," *Universidade Federal do Rio de Janeiro, Rio de Janeiro*, 2017.

[36] C. M. Bautista, C. A. Dy, M. I. Mañalac, R. A. Orbe, M. Cordel, "Convolutional neural network for vehicle detection in low-resolution traffic videos," in *2016 IEEE Region 10 Symposium (TENSYMP)*, 2016, pp. 277–281, IEEE.

[37] X. Zhao, Y. Chen, M. Tang, J. Wang, "Joint background reconstruction and foreground segmentation via a two-stage convolutional neural network," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 343–348, IEEE.

[38] J. Wang, K. L. Chan, "Background subtraction based on encoder-decoder structured cnn," in *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, 2020, pp. 351–361, Springer.

[39] Y. Wang, Z. Luo, P.-M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognition Letters*, vol. 96, pp. 66–75, 2017.

[40] Y. Zhang, D. Zhou, S. Chen, S. Gao, 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.

[41] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] L. Boominathan, S. S. Kruthiventi, R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 640– 644.

[43] Y. Li, X. Zhang, D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[44] G. Gao, J. Gao, Q. Liu, Q. Wang, Y. Wang, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.

[45] V. Lempitsky, A. Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010.

[46] S. Kumagai, K. Hotta, T. Kurita, "Mixture of counting cnns," *Machine Vision and Applications*, vol. 29, no. 7, pp. 1119–1126, 2018.

[47] Y. Zhang, S. J. Kiselewich, W. A. Bauson, R. Hammoud, "Robust moving object detection at distance in the visible spectrum and beyond using a moving camera," in *2006 conference on computer vision and pattern recognition workshop (CVPRW'06)*, 2006, pp. 131–131, IEEE.

[48] K. K. Verma, B. M. Singh, "Deep multi-model fusion for human activity recognition using evolutionary algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 44– 58, 2021.

[49] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, S. Yan, "Crowded scene analysis: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 25, no. 3, pp. 367–386, 2014.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[51] L. Liu, J. Jiang, W. Jia, S. Amirgholipour, Y. Wang, M. Zeibots, X. He, "Denet: A universal network for counting crowd with varying densities and scales," *IEEE Transactions on Multimedia*, vol. 23, pp. 1060–1068, 2020.

[52] L. Liu, H. Lu, H. Zou, H. Xiong, Z. Cao, C. Shen, "Weighing counts: Sequential crowd counting by reinforcement learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 2020, pp. 164–181, Springer.

[53] M. Hossain, M. Hosseinzadeh, O. Chanda, Y. Wang, "Crowd counting using scale-aware attention networks," in *2019 IEEE winter conference on applications of computer vision (WACV)*, 2019, pp. 1280–1288, IEEE.

[54] X. Ding, Z. Lin, F. He, Y. Wang, Y. Huang, "A deeply- recursive convolutional network for crowd counting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1942–1946, IEEE.

[55] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, A. G. Hauptmann, "Learning spatial awareness to improve crowd counting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6152–6161.

[56] H. Idrees, I. Saleemi, C. Seibert, M. Shah, "Multi- source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[57] A. B. Chan, Z.-S. J. Liang, N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE conference on computer vision and pattern recognition*, 2008, pp. 1–7, IEEE.

Ankit Tomar

Mr. Ankit Tomar has done B Tech Computer Science and Engineering from UPTU, M-Tech from Jamia Hamdard, New Delhi, and pursuing a Ph.D. from Graphic Era Deemed University, Dehradun. He is actively involved in research related to Deep Learning and Machine Learning. He has published many papers in reputed international conferences and journals. Currently, he is working as an assistant professor in the Graphic era deemed by the university Dehradun, India.

Santosh Kumar

Dr. Santosh Kumar earned his Ph.D. from India Institute of Technology, Roorkee, India in 2012, M. Tech. in Computer Science and Engineering from Aligarh Muslim University, Aligarh, India in 2007, and B.E. (IT) from C.C.S. University, Meerut, India in 2003. He is an active reviewer board member in various national/International Journals and Conferences. He has memberships of ACM (Senior Member), IAENG, ACEEE, and ISOC (USA) and contributed more than 80 research papers in National and International Journals/conferences. Currently holding a position of Professor in the Graphic Era Deemed to be University, Dehradun, India. His research interest includes AI &; Machine Learning, Wireless Networks, WSN, IoT, and Software Engineering.

Bhaskar Pant

Dr. Bhaskar Pant is currently working as Dean of Research & Development and Associate Professor in the Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal. Has more than 17 years of experience in Research and Academics. He has till now guided as Supervisor 4 Ph.D. candidates (Awarded).and 7 candidates are in an advanced state of work. He has also guided 32 M. Tech. Students for dissertation. He has also supervised 2 foreign students for internship. Dr. Bhasker Pant has more than 100 research publications in National and international Journals.