Crowd Counting and Localization Beyond Density Map

AKBAR KHAN1, KUSHSAIRY KADIR1, (Senior Member, IEEE), HAIWATI NASIR2, (Senior Member, IEEE), JAWAD ALI SHAH3, (Senior Member, IEEE), WALEED ALBATTAH4, (Senior Member, IEEE), SHEROZ KHAN5, (Senior Member, IEEE), AND MUHAMMAD HARIS KAKAKHEL1

1Electrical Section, Universiti Kuala Lumpur British Malaysian Institute, Selangor 53100, Malaysia
2Department of Electrical Engineering, International Islamic University, Islamabad 44000, Pakistan
3Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
4Department of Electrical Engineering, College of Engineering and Information Technology, Qassim 52053, Saudi Arabia
5Department of Electrical Engineering, Universiti Kuala Lumpur Malaysian Institute of Technology, Kuala Lumpur 50250, Malaysia

Corresponding author: Kushsairy Abdul Kadir (kushsairy@unikl.edu.my)

This work was supported by the UniKL Publication Support Grant.

ABSTRACT Crowd analysis in general and counting in congested scenes, in particular, is an effective and vibrant research domain in computer vision due to its numerous applications. Understanding the risk analysis and safety aspects of crowd dynamics at various vital occasions related to sports cultural and religious activities, specifically, at Hajj and Umrah, is essential. Thousands of people gathered in a small area to carry out their rites. Localizing and counting the annotated head points is quite challenging due to occlusion and large-scale variation in the congested environment. To deal with these problems, a small and effective solution is to generate the density maps. However, the significant flaws of the density map have a blurry Gaussian blob which is less effective for counting and localizing head annotations in the congested scene. To overcome these issues, we propose Congested Scene Crowd Counting and Localization Network (CSCCL-Net) with a Focal inverse Distance Transform (FIDT) map that can count and localize the people simultaneously in the highly congested scene. To evaluate the proposed model’s efficiency, extensive tests were performed on the ShanghaiTech part A, ShanghaiTech part B, and JHU-CROWD++ datasets. The proposed model outperforms existing state-of-the-art techniques regarding high accuracy and low Mean Absolute Error (MAE) and Mean Square Error (MSE) values.

INDEX TERMS Crowd counting, crowd localization, deep learning, focal inverse distance transform, convolutional neural networks.

I. INTRODUCTION

Object counting and localization are evolving research nowadays in the area of computer vision. Both crowd counting and localization have drawn attention interesting from the research community. It shows a significant role in maintaining public safety and security at various events like traffic management, urban planning, monitoring of smart cities, political rallies, and religious gatherings of Hajj and Umrah. Since crowd counting and localization are closely related to each other, therefore, it is necessary to describe them briefly. The primary aims of counting are to estimate the objects in images or videos. In contrast, localization means predicting the concise location of that object in the images or videos. The crowd monitoring and management in the congested scene are crucial due to the recent Covid-19 pandemic across the globe. The counting and localization of people in congested scenes encounter many key challenges. It can be scale variation, overlapping with severe occlusion and due to blurry background [1], [2], [3], [4], [5]. Mainly, there are three common approaches generally used for crowd counting, that is, detection-based crowd counting [6], [7], [8], [9], regression-based crowd counting, and density-based crowd counting [10], [11]. In detection-based
crowd counting, a sliding window is used to detect and count the number of people in an image. The disadvantage of this approach is that it cannot detect and count people in a very crowded scene [12]. Regression-based approaches have been used to regress the low-level feature information to overcome this drawback. This approach can do well in a highly congested scene because of the collection of generalized density information. The regression-based approaches are unable to regress in occlusion and cluttered areas. To some extent, the density-based approaches can address these challenges but possess various challenges, that is, blurry gaussian blobs and severe overlapping with nearest head annotations, as shown in Figure 1. While for crowd localization, the authors have explained three major methods: detection-based crowd localization, heuristic-based crowd localization, and point-supervision crowd localization [13], [14]. Object detection is a valuable method for crowd counting and localization [15]. As count signifies the number of detected bounding boxes in the image, however, the drawbacks of the detection method are that it can suffer from severe occlusion and overlapping of bounding boxes. Additionally, it is quite difficult to annotate the bounding boxes in the congested scenes due to being expensive and laborious. The concept of density maps was proposed to handle these challenges in [16], [17], [18], [19], and [20], which is formed by the Gaussian kernel. Convolutional Neural Networks (CNNs) are used in several counting approaches to regress density maps with good results. However, these methods [17], [18], [21], [22], [23] have failed to anticipate the precise location of each head in the congested scene. And even these methods cannot count and locate the objects in an image or video simultaneously due to a series of blurry Gaussian blobs and overlapping. Because of the above mentioned issues, we propose CSCCL-Net, which can simultaneously count and localize the people in the congested scene. FIDT map is used to pinpoint the exact position of people. At the same time, it is evident that the density map depicts blurry Gaussian blobs in a crowded area. It is also unrecognizable owing to the Gaussian kernel filtering on each head annotation. Even then, in an FIDT map, every point is visible and distinguishable even in the congested scene, as mentioned earlier. The Local Maxima Detection Strategy (LMDS) [25] finds the head coordinates to pinpoint the local maxima and exhibits flexibility to negative background data. K-nearest neighbor is used to building bounding boxes for each head. Meanwhile, before training, the bounding boxes were produced using ground truth dotted annotation in [26] and [27]. We have generated the bounding boxes before training through predicted positions. Counting by localization requires comprehensive structural information in the FIDT map at the local level. The SSIM loss can be used to increase the resemblance between the predicted FIDT map and the ground truth map. The conventional SSIM loss may result in high background responses that are unstable and create false local maxima, amplifying the localization and counting error. Then, we utilized an independent SSIM loss to improve the model’s ability to increase structural information of local maxima while decreasing false local maxima in the background. VGG-16 has been used as a core network and its first ten layers as a front-end in the proposed framework. A hole dilated convolutional neural network has been employed as a back-end to enlarge the respective fields to get more information from the image and decrease the false local maxima in the background area.

As a result of the issues mentioned above, it is essential to solve them in the suggested model. In summary, the contributions are as follow:

1. Proposed a practical model that can cope with both counting and localization tasks at the same time, based on a congested scene recognition network and the focal inverse distance transform
2. The FIDT map has been enhanced to improve the blurry Gaussian blobs
3. In addition, The proposed model has deduced the findings on ShanghaiTech part A, B, and JHU-CROWD++ datasets, respectively

The structure of the article includes: Section II summarizes previous work on crowd counting and localization research. In Section 3, we presented a crowd counting and localization architecture. The experiment findings are evaluated in Section 4, followed by results and discussion in Section 5. Finally, Section 6 summarises the conclusions.

II. PREVIOUS WORK

During the past decade, deep learning made significant progress in computer vision. Several deep learning approaches for image segmentation, classification, and detection have been developed with significant success in the literature. Different CNNs and deep CNNs with multi view models have been published in the literature to count and find individuals in images and videos, motivated by the considerable success of deep learning [2], [28].
A. CROWD COUNTING

Various prominent deep learning crowd counting models are included in this section. The three types of deep learning crowd counting models are: 1) detection-based crowd counting, 2) regression-based crowd counting, and 3) density-based crowd counting [2], [10], [11], [18].

1) DETECTION-BASED CROWD COUNTING

Throughout the early phases of deep learning, researchers concentrated on detection-based methods for crowd counting. Sliding windows are employed in these approaches to detect, count, and localize the objects in images or videos [10], [13], [15], [29]. To extract low-level features from the whole human body, many methods, such as Haar wavelets [30] and Histogram of Oriented Gradients (HOG) [15] are reported, which require well-trained classifiers. There are two common styles used in detection-based crowd counting, namely:

1. Monolithic style
2. Part-based detection style

Earlier, a classifier was trained employing features taken from the entire body. In the latter case, the classifier is trained to recognize specific points, such as the head, for example to estimate the number of individuals. The authors in [31] have suggested a technique for calculating the number of individuals using MID-based foreground segmentation and heads detection to estimate the number of people. These two sub-methods are unable to detect and count the number of people in the highly congested scenes.

2) REGRESSION-BASED CROWD COUNTING

Considering the above two issues mentioned in detection-based methods are not applicable in the highly congested scenes. Therefore regression-based crowd counting has been introduced to overcome these issues [2], [32]. This approach contains two steps of low-level feature extraction and modeling of regression. Image features and crowd size are utilized to estimate the crowd count using regression-based algorithms [17], [32], [33], [34], [35], [36]. The features that can be extracted from the image patches are:

- Texture features
- Edge features
- Gradient features
- Foreground features

Low-level data is generated using these attributes. For feature extraction, various regressing techniques are used, including Gaussian process regression, linear regression, and piece wise linear regression [10]. Likewise in [32], a Fourier series has been employed to extract the feature from the image. Similarly, [17] uses a Multi-Column Convolutional Neural Network that has employed three columns with different filter sizes to compensate for it. CNN models with two configurations estimate the crowd density in a single image [16]. While switched CNN uses multiple CNN for crowd counting [21]. In addition, some models are based on attention mechanisms to cope with the crowd counting via regression-based [37], [38], [39]. Specifically, the authors in [37] have developed local and global self-attention to capture short-range and long-range information. ADCrowdNet uses a binary classification network for the regression of the crowd counting region [39]. Similarly, density maps were used as probability maps to get the probability of each pixel in the image [33]. The Distribution Matching (DM) approach compares a normalized projected density map to a normalized ground truth density map using optimum transport [34]. Kernel-Based Density Map Generation (KDMG) is a learnable density map representation that uses an adaptive density map generator. They have focused on counting rather than the process of localization [40]. However, the performance of regression-based still suffers from occlusion and scale variation.

3) DENSITY-BASED CROWD COUNTING

Since the effectiveness of crowd counting has improved dramatically when spatial information is integrated, density-based crowd counting has become increasingly popular. Dot maps, with each dot representing a person in the image, are typically blurred to make density maps [40]. Most approaches construct density maps in advance by convolving dot maps with Gaussian kernels with fixed or adjustable bandwidths. Then, to solve various barriers, such as size fluctuation, improving density map quality, encoding more contextual information, or adapting to changing conditions, alternative network models are built. To accommodate scale fluctuation, the Multi-Column Neural Network (MCNN) proposes extracting features from many columns with various Kernel sizes [17], [19]. Switch-CNN suggests selecting a right column with an adequate receptive field rather than combining multi-scale features. To manage the diversity of individuals in crowds, a tree-structured CNN has been proposed by authors
B. CROWD LOCALIZATION

Recently, crowd localization has been a new and active research topic in computer vision. It takes the individual as a basic unit instead of the scenes [13]. The research community has paid less attention to gathering evidence regarding crowd localization in densely populated areas. Finding the person’s exact position in a crowd is a challenge in localization. The distribution of individuals in the environment may be determined via localization, which is critical for crowd managers to maintain the safety and security of the public. Furthermore, it may be utilized in a dense crowd to recognize and track a person. And to generate ground truth data, which can be used to correct counting errors [2]. To locate the head position of each individual in the congested crowd scene, some researchers have focused on crowd localization, consisting of three types, namely: 1) detection-based crowd localization, 2) point-based, and 3) heuristic-based crowd localization [14].

1) DETECTION-BASED CROWD LOCALIZATION

There are various ways for locating individuals heads in a crowd that employs bounding boxes. Methods based on detection train object detectors to pinpoint each person’s location [29], [43], [44]. In [45] and [46], the authors have proposed head detection method using hand-crafted features for a real-time surveillance system. As hand-crafted features are limited, therefore, several CNN techniques are used to enhance the performance of complex scenes. Similarly, the authors in [47] present context-aware head detection method for crowd localization. A recurrent LSTM technique has been built for sequence creation, and its detector has been used to obtain the detection results [6].

2) POINT-BASED CROWD LOCALIZATION

The most of the crowd datasets contain images which annotated by points instead of bounding box annotations. As a result, it is more practical and useful to employ point annotations for a crowd as bounding box annotation as result productive. After recasting the crowd localization problem as a foreground/background segmentation problem, they used the cross-entropy loss to optimize the network [23], [48], [49], [50].

3) HEURISTIC-BASED CROWD LOCALIZATION

The heuristic-based crowd localization is used to extract the crowd location from density maps proposed in [51] and [52]. Non-maximum suppression has been used to obtain the highest local value. This shows the head annotations of each individual. After that 1-1 matching has been adopted to match the true head location. Finally, the feasible possible solution was obtained by applying the Hungarian algorithm for the evolution of crowd localization.

Apart from the above-proposed methods, we have used focal inverse distance transform that can count and localize the annotated heads in the congested scenes. Compared to density maps, Focal Inverse Distance Transform (FIDT) maps appropriately indicate a person’s location in a congested scene without overlapping heads. We regress the FIDT maps at the same time to determine the person’s position and the count.

C. LIMITATIONS

Existing approaches mainly rely on crowd counts while employing a density map. The primary disadvantage of a density map is that it is difficult to detect the specific position of individuals in dense images due to blurry Gaussian blobs. These methods seldom pay attention to the particular placement of persons in images in crowded settings. This leads our study toward developing a viable combined solution for crowd counts and localization based on the FIDT map.

III. PROPOSED ARCHITECTURE

The proposed architecture is depicted in Figure 3 in this section. Random cropping and horizontal flipping are used to enhance the input training images at first. We used data augmentation techniques to boost efficiency and minimize overfitting because most crowd-counting datasets are small in size. At first, we execute random cropping and resizing the training set with different sizes to overcome the problem of scale variation and retain performance on zoomed-in images. Then these cropped images have been used to generate ground truths. The cropped images’ size for ShanghaiTech part A dataset and part B dataset are 256 × 256, and the datasets have images’ size of 512 × 512. A regressor is used to regress the predicted FIDT maps. We have used a congested scene recognition network with different dilation rates as the core network to train the model.

A. MATHEMATICAL ELABORATION OF FOCAL INVERSE DISTANCE TRANSFORM MAP

Many deep learning-based approaches use FIDT map as an image processing tool [25], [53]. The FIDT map is used to count and locate the crowds simultaneously without overlapping the heads in congested scene. It is mathematically expressed in (1):

\[
P(x, y) = \min_{(x', y') \in B} \sqrt{(x - x')^2 + (y - y')^2}
\]
where \( B \) is the total number of annotations for an arbitrary pixel \( P(x, y) \). The equation (1) represents the distance between a pixel and its nearest head annotation. Due to the high distance variations, the distance transform map is not easy to regress immediately. An Inverse Distance Transform map (IDT) has been used to prevent variations in distance, as defined by (2):

\[
I' = \frac{1}{P(x, y) + C}
\]

where \( I' \) represents the IDT map, and \( C = 1 \). The IDT map is a unique form of iKNN map. Instead of crowd localization, the iKNN map is employed for crowd counts. The IDT (iKNN) map on the other hand, may properly depict specific places that correlate to local maxima when compared to frequently used density maps.

The response time should be slower while going away from the head, and the background should go close to 0 as fast as possible. The foregrounds should be emphasized by the head, and the background should go close to 0 as fast when compared to frequently used density maps.

As a result, the Focal Inverse Distance Transform (FIDT) is proposed. The mathematical expression of FIDT map is given in (3):

\[
I = \frac{1}{P(x, y)^{(\alpha \times P(x,y) + \beta)} + C}
\]

where \( I \) represents the FIDT map and \( \alpha, \beta \) are the statistical parameters used for contrast and illumination, respectively.

### B. THE EFFECTIVENESS OF FIDT MAPS FOR CROWD LOCALIZATION

The HRNET for regression was used to assess the efficiency of the FIDT maps. High-resolution representations are required for position-sensitive vision issues such as human posture estimation, semantic segmentation, and object detection [54]. The counting results of the proposed FIDT map are similar to local maxima. Therefore, high representation is required. We can locate the person’s position using a FIDT map with local maxima.

### C. ELABORATION OF LOCAL MAXIMA DETECTION STRATEGY

The position of the person in the highly congested scene can be easily found via a predicted FIDT map. In the Local Maxima detection Strategy (LMDS), we first use the \( 3 \times 3 \) max-pooling methods to get candidate points. However, some false positives may exist in the background of these candidate points. We see those true positives have a significantly higher pixel value than false positives, meaning that a local maximum is prone to exist. As a result, it has been decided to use threshold values to eliminate false positives. Hence, given a set of candidate points \( M \), the last chosen points are those whose value is more than or equal to an adaptive threshold, equal to \( 100/255.0 \) times \( M \)'s maximum value. The most recent dataset [55] contains some negative samples comparable to crowd scenes and some scenes without humans. Based on the anticipated density maps, we can not tell if the original images have people in them. Given a predicted FIDT map, if the maximum of \( M \) is less than a small number (set at 0.10), this indicates that the input image is a negative sample, and the LMDS will put the counting result to 0. We can build pseudo individual sizes using the proposed FIDT map, even when actual individual sizes are unavailable. We first extract the coordinates of the head centers from the predicted FIDT map, which can be effectively implemented using the suggested LMDS. The instance size is then calculated by applying the K-nearest neighbor distance, as defined:

\[
S_{(x,y)\in P} = \min \left\{ \bar{d} = f \times \frac{1}{k} \sum_{k=1}^{k} d_{(x,y)}^k \right\}
\]

where \( S_{(x,y)} \) show off the size of the instance on the x-axis and y-axis, respectively with \( P \), i.e., a collection of predicted head points. The \( \bar{d} \) is the average distance calculated between \( P(x,y) \) and its k-nearest neighbors. \( f \) are used to confine the size of the predicted FIDT map. In particularly sparse locations, the \( \bar{d} \) maybe larger than the actual size of individuals. Thus we pick and to limit the object size, choose a threshold based on image size.

### IV. EXPERIMENTAL EVALUATION

The experiments are conducted on different mainstream datasets, namely, ShanghaiTech parts A, B, and JHU-CROWD++. The latest dataset, i.e., JHU-CROWD++ is the most challenging for crowd counting and localization owing to its several illumination variations and complex background. The key information of the datasets is listed in Table 1.

#### A. IMPLEMENTATION DETAILS

In implementation details, we augment the datasets with random cropping and horizontal flipping with a size of \( 256 \times 256 \) for ShanghaiTech parts A and B, and \( 512 \times 512 \) for other datasets, such as JHU-CROWD++, which have a diversified scene. To produce the bounding boxes, we set \( k \) to 4 and \( f \) to 0.1 further for optimization of the model Adam [56] has been used, with a learning rate of \( le - 4 \) and a weight decay of \( 5 \times le - 4 \). For threshold selection, the distance between the given predicted point \( P_p \) and ground truth point \( P_g \) must be less than a distance threshold \( \alpha \) (real head size). It indicates that \( P_p \) and \( P_g \) have been successfully matched. The MSELoss has been used as a loss function with addition of Adam as a optimizer throughout the training phases.

#### B. EVALUATION PROTOCOL

This section uses the Mean Absolute Error (MAE) and Mean Squared Error (MSE) as statistical measures to evaluate counting performance. These statistical parameters are being defined mathematically as:

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|
\]
TABLE 1. Details of datasets.

| Dataset          | ShanghaiTech part A [17] | ShanghaiTech part B [17] | JHU-CROWD++ [57] |
|------------------|---------------------------|---------------------------|-------------------|
| Number of images | 482                       | 716                       | 4372              |
| Training images  | 300                       | 400                       | 2272              |
| Testing images   | 182                       | 316                       | 1600              |
| Number of annotations | 330,165                  | 15,15,005                 |                   |
| Max Count        | 3,139                     | 578                       | 25791             |
| Average HxW      | 589 x 868                 | 768x1024                  | 910x1430          |

\[
\text{MSE} = \left( \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|^2 \right)^{1/2} \tag{6}
\]

We examined the crowd localization performance on multiple statistical criteria, such as Precision, Recall, and F1 measure, and crowd counting metrics. They are mathematically defined as:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{7}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{8}
\]

\[
\text{F1 measure} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{9}
\]

C. DATASETS

We evaluate the approach on three popular datasets, each discussed in detail below.

1) ShanghaiTech DATASETS

[17] proposed the ShanghaiTech datasets, which are divided into two parts: part A and part B. Part A has 482 images and 241,677 instances, whereas part B has 716 images and 88,488 labeled heads.

2) JHU-CROWD++

There are 4,372 images with 1.51 million annotations in this dataset. Unlike existing datasets, the proposed dataset was obtained in various scenarios and environments. In addition, the dataset includes a broader range of annotations, such as dots, approximate bounding boxes, and blur levels. It consists of 2722 training, 500 validation, and 1600 test images. It contains diverse crowd scenes, such as fog, rain, snow, and low illumination [57].

D. COMPARISON WITH STATE-OF-THE-ART-METHODS

We compared our proposed framework with other state-of-the-art methods like [21], [58], [59], and [60] on ShanghaiTech part A, and part B datasets. We report the results of these latest existing methods, shown in Table 2. From Table 2, it is demonstrated that our proposed methods compete and outperform the recent existing methods. Reference [61] has proposed a deep fusion network that combines multi-scale features from shallow to deep network. Peaks of density maps achieved the final count. But the drawback of density maps is blurry gaussian blobs specifically, it is challenging in the highly congested scene with the problem of overlapping.

V. RESULTS AND DISCUSSION

We have tested and assessed our proposed model on various datasets. The training and validation losses has been shown in Figure 4. The experimental results on ShanghaiTech part A and Part B datasets of the proposed model have been compared with recent state-of-the-art methods based on crowd counting and localization. In Table 2 and Table 3, it is demonstrated that the proposed model has achieved remarkable performance in terms of MAE, MSE, average precision, average recall, and F1 score. For ShanghaiTech Part A and Part B datasets as listed in Table 2 the proposed model reports SOTA counting performance compared with recent models [14], [59], [61]. Similarly, the proposed model has been compared with recent existing methods for crowd localization and achieved notable results as shown in Table 3. For crowd localization, the proposed model can identify the exact position.
A. JHU-CROWD++ BENCHMARKING

Crowd counting datasets have changed over time in terms of size, crowd density, image resolution, and various other factors. The UCSD dataset has been created early and consists of 2000 low-resolution video frames with 49,885 annotations [70]. Similarly, the UCF_CC_50 with 50 images has been proposed [32] based on high-density crowd images. But the drawback of this dataset is less number of Images, which is not reliable for training the deep learning networks. The ShanghaiTech dataset has been introduced by [17] for better variety in terms of both scene and density levels.

1) EVALUATION OF COUNTING AND LOCALIZATION

First, we compare the proposed approach to existing and cutting-edge crowd counting and localization methods. The qualitative results on JHU-CROWD++ dataset have been shown in Figure 5. It has been stated that the existing models are based on both crowd counting and localization simultaneously. Based on the results, we can see from Table 4, and Table 5 that our model has achieved good performance in terms of both MAE and MSE and produces the correct counting task. The CSCCL-Net plus FIDT model improves the counting and localization performance in highly dense crowd images, including fog, rain, snow, low illumination, and so on. A multi-scale contextual approach has been proposed in [71] that combines the features obtained from the multiple respective fields to predict the crowd density. An encoder and decoder network has also been used to extract multi-scale features using a set of transposed networks. Further, the model’s efficiency has been improved by the combination of Euclidean loss and local pattern consistency loss [19]. To leverage the information presented in different layers of the network, a multi-scale fusion has been used to improve the model’s effectiveness [42]. The point annotations create a density contribution probability model using Bayesian loss. The predicted count at each annotated location is then determined by adding the contribution probability and estimated density at each pixel, which is reliably guided by the ground truth count value [33]. A crowd counting network with a rich set of annotations at image and head levels progressively generates the density maps. The proposed model has used VGG-16 as a core network with the addition of generated density map for the final prediction of the people in the crowd [57]. An adaptive density map generator that accepts an annotated dot map as input and learns a density map representation for a counter has also been presented for crowd counting [40]. As shown in Table 4, and Table 5, we compared and assessed our model’s performance for crowd counting with recent existing SOTA methods. The graphical representation of JHU-CROWD++ dataset has shown in Figure 6. On the JHU-CROWD++ dataset, the suggested model has steadily reduced the MAE and MSE, and the findings show that the proposed model has achieved exceptionally good numerical results.

VI. ABLATION STUDY

We evaluate the performance of our proposed model in relation to the backbone network with the existing baseline model. The backbone network in the proposed model is the VGG-16. We conduct the experiments using the JHU-CROWD++ dataset, and the results clearly show that it...
performs better than the current standard. Table 5 displays the efficacy of the suggested model with higher performance in terms of average accuracy of 5% and F1 measure of 3% accordingly.

VII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.

X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, “Repulsion loss: Detecting pedestrians in a crowd,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7774–7783.

V. Conclusions and Future Work
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.



VIII. CONCLUSION AND FUTURE WORK
This paper has proposed a congested scene crowd counting and localization simultaneously using VGG-16 as the core network with the addition of dilated convolutional neural network. The proposed framework is based on point-level annotations. Notably, using the FIDT map can remove the blurry Gaussian blobs in the density map and promptly enhance the performance. Extensive experiments demonstrate that this framework has achieved notable crowd counting and localization results in the case of highly congested scenarios, including diverse environments. As shown in Table 4, our proposed CSCCL.Net has achieved MAE 61.5 and MSE 226.2 compared to other networks. Equivalently, our proposed CSCCL.Net has achieved notable crowd counting and localization using deep convolutional neural network: A Review, Appl. Sci., vol. 10, no. 14, pp. 4781, Jul. 2020.
C. Liu, X. Weng, and Y. Mu, “Recurrent attentive zooming for joint crowd counting in pedestrian crowds,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1217–1226.

A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 270–276.

Y. Ishii, H. Hongo, K. Yamamoto, and Y. Niwa, “Face and head detection for a real-time surveillance system,” in Proc. 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 298–301.

S. D. Khan and S. M. Basalamah, “Sparse to dense scale prediction for crowd counting in high density crowds,” Arab. J. Sci. Eng., vol. 46, no. 4, pp. 3051–3065, 2020.

Y. Shi, P. Mettes, and C. Snoek, “Counting with focus for free,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4200–4209.

P. Hu and D. Ramanan, “Finding tiny faces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 951–959.

A. Zhang, L. Yue, J. Shen, F. Zhu, X. Zhen, X. Cao, and L. Shao, “Attentional neural fields for crowd counting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 5714–5723.

S. D. Khan, S. M. Basalamah, “Sparse to dense scale prediction for crowd counting in high density crowds,” in Proc. 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 298–301.

S. D. Khan and S. M. Basalamah, “Sparse to dense scale prediction for crowd counting in high density crowds,” in Proc. 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 298–301.

T.-H. Vu, A. Osoin, and I. Laptev, “Context-aware CNNs for person head detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2893–2901.

Z. Shi, P. Mettes, and C. Snoek, “Counting with focus for free,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4200–4209.

N. Pravc, A. Sustar, M. Perkovic, G. Erjavec, and D. Struc, “Association between reduced brain-derived neurotrophic factor concentration & coronary heart disease,” Indian J. Med. Res., vol. 150, no. 1, p. 43, 2019.

C. Xu, D. Liang, Y. Xu, S. Bai, W. Zhan, X. Bai, and M. Tomizuka, “AutoScale: Learning to scale for crowd counting and localization,” 2019, arXiv:1912.09632.

H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, “Composition loss for counting, density map estimation and localization in dense crowds,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 532–546.

S. D. Khan, H. Ullah, M. Uzair, M. Ullah, R. Ullah, and F. A. Cheikh, “DISAM: Depth independent and scale aware model for crowd counting and localization,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 4474–4478.
AKBAR KHAN received the B.E. degree (Hons.) in computer system engineering from the Balochistan University of Engineering and Technology (BUET) Khudzdar, in 2011, and the M.S. degree in computer engineering from the Balochistan University of Information Technology Engineering and Management Sciences (BUTEEMS), Quetta, Pakistan, in 2017. He is currently pursuing the Ph.D. degree with the Universiti Kuala Lumpur British Malaysian Institute, Malaysia. He joined the Department of Computer Engineering, BUTEEMS, as a Lecturer, in 2013, where he is currently working as an Assistant Professor. His research interests include machine learning and computer vision.

KUSHAIRY KADIR (Senior Member, IEEE) received the Bachelor of Science degree in engineering from the University of the West of England, Bristol, in 1998, the Master of Science degree in mechatronic from International Islamic University, in 2007, and the Ph.D. degree in electronic and electrical engineering from the University of Strathclyde, in 2012. He is currently an Associate Professor with the Electrical Technology Section, Universiti Kuala Lumpur British Malaysian Institute and the Dean of the UniKL British Malaysian Institute. His research interests include signal image processing, robotics for rehabilitation, and building energy efficiency. He is also the Deputy Chair of the IEEE IMS Malaysia Chapter.

HAIDAWATI NASIR (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Electronic, University of Strathclyde, Glasgow, in 2012. She is currently an Associate Professor with the Computer Engineering Section, Universiti Kuala Lumpur Malaysian Institute of Information Technology. Her research interests include signal and image processing with application to super-resolution image reconstruction, image enhancement, and computer vision. She is also a Student Activities Chair of IEEE Malaysia Section and a Treasurer of the IEEE SPS Malaysia Chapter.

JAWAD ALI SHAH (Senior Member, IEEE) received the B.S. degree (Hons.) in electrical engineering and the M.S. degree (Hons.) in telecommunication engineering from the University of Engineering and Technology Peshawar, Pakistan, in 2001 and 2007, respectively, and the Ph.D. degree from International Islamic University, Islamabad, Pakistan, in 2015. He is currently an Assistant Professor with the Department of Electrical Engineering, International Islamic University. His current research interests include sparse signal processing, compressed sensing, and machine learning. He won two gold medals and presidential award in his academic career. He was a recipient of a six-month fellowship by Higher Education Commission Pakistan under the International Research Support Initiative Program (IRSSIP) to pursue his research work at UC Denver/Anschutz Medical Campus, Aurora, CO, USA.

WALEED ALBATTAH (Senior Member, IEEE) received the Ph.D. degree from Kent State University, Kent, OH, USA. He is currently a Faculty Member with the Department of Information Technology, Qassim University, Saudi Arabia. He is also working on big data and cloud computing security as well as machine learning projects. He has been granted several project grants from different organizations. His research interests include software engineering, software design and agile software development, and software quality. He is a member of the ACM Society SIGSOFT.

SHEROZ KHAN (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the NWFP University of Engineering and Technology (UET) Peshawar, Pakistan, the M.Sc. degree in microelectronic and computer engineering from the University of Surrey, U.K., in 1988, and the Ph.D. degree in electrical and electronic engineering from the University of Strathclyde, U.K. He rejoined his parent university, in August 1994. After serving the NWFP UET for five years, he was selected by a Malaysian delegation for working as a Principal Lecturer at the UNITEN, since January 2000. Upon completing his two-year contract at UNITEN, he started as an Associate Professor at International Islamic University Malaysia, since January 2002. He has been the Postgraduate Co-ordinator with the Department of ECE, IIUM, and the Founding Co-ordinator of the Wireless Communication and Signal Processing Research Group, since 2006. He has been the Founder of the IIUM-Limoges (France), IIUM-Schmalkalden UAS (Germany), and IIUM-IIU Islamabad (Pakistan) link programs. He has been working with the Department of Electrical Engineering, Onaizah College of Engineering and Information Technology, since December 2019, where is charged with the task of organizing an international conference. He has so far produced 21, M.Sc. under his direct supervision and ten M.Sc. under co-supervision. He has produced nine Ph.D. and two Ph.D. under his direct supervision while producing six Ph.D. under co-supervision. He is the Founding Counselor of the IEEE Student Branch (STB64633). Being the Best Graduate of the Department of Electrical Engineering, in 1982, he was awarded the University Scholarship for doing M.Sc. degree.

MUHAMMAD HARIS KAKAKHEL was born in Charsadda, Khyber Pakhtunkhwa, Pakistan, in January 1997. He received the B.S. degree in electrical engineering and communication from the University of Engineering and Technology (UET) Peshawar, Pakistan, in 2019. He is currently pursuing the master’s degree with the Universiti Kuala Lumpur British Malaysian Institute, Malaysia, under the supervision of Dr. Kushsairy Abdul Kadir. His research interests include computer vision, deep learning, image processing, and wireless communications.