IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 25, 2019, accepted December 11, 2019, date of publication December 17, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960292

CASA-Crowd: A Context-Aware Scale Aggregation CNN-Based Crowd Counting Technique

NAVEED ILYAS[1], ASHFAQ AHMAD[2], (Student Member, IEEE), AND KISEON KIM[1], (Senior Member, IEEE)

[1]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea
[2]School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia

Corresponding author: Ashfaq Ahmad (ashfaqahmad@ieee.org)

This work was a part of project titled, Development of Ocean Acoustic Echo Sounders and Hydro-Physical Properties Monitoring System, funded by the Ministry of Oceans and Fisheries, South Korea.

ABSTRACT The accuracy of object-based computer vision techniques declines due to major challenges originating from large scale variation, varying shape, perspective variation, and lack of side information. To handle these challenges most of the crowd counting methods use multi-columns (restrict themselves to a set of specific density scenes), deploying a deeper and multi-networks for density estimation. However, these techniques suffer a lot of drawbacks such as extraction of identical features from multi-column, computationally complex architecture, overestimate the density estimation in sparse areas, underestimating in dense areas and averaging of feature maps result in reduced quality of density map. To overcome these drawbacks and to provide a state-of-the-art counting accuracy with comparable computational cost, we therefore propose a deeper and wider network: a Context-aware Scale Aggregation CNN-based Crowd Counting method (CASA-Crowd) to obtain the deep, varying scale and perspective varying features. Further, we include a dilated convolution with varying filter size to obtain contextual information. In addition, due to different dilation rates, a variation in receptive field size is more useful to overcome the perspective distortion. The quality of density map is enhanced while preserving the spatial dimension by obtaining a comparable computational complexity. We further evaluate our method on three well-known datasets: UCF_CC_50, ShanghaiTech Part_A, ShanghaiTech Part_B.

INDEX TERMS Deep learning, convolutional neural networks, density estimation, crowd counting.

I. INTRODUCTION

Automated crowd counting refers to estimating the number of individuals in unconstrained scenes depicted by images and videos. It has applications in crowd management, urban planning, congestion avoidance, flow analysis, anomaly detection, video supervenience, public safety management, defense, health-care, disaster management, so on. Nonetheless, it is very challenging to accurately obtain the count due to severe occlusion, clutter, irregular object distribution and non-uniform object scale [1]–[4].

The number of people and their spatial distribution are two significant measurements for understanding crowded scenes. Detection, tracking and counting in low resolution images and surveillance videos, where people are represented by only few pixels tall, are issues that yet demand more investigation. Detection-based crowd counting performs well in sparse crowd, however, the performance of both detection and counting degrades with increase in object density. In contrast, regression-based crowd counting methods [1], [2] perform well in very dense crowded environment. However, it faces significant hurdles while adapting the scale variation and preserving the spatial information. With the boost of convolutional neural networks (CNN), various CNN-based crowd counting algorithms have mushroomed for addressing the difficulties of crowd counting. One of the most significant advantages of CNN-based crowd counting is its ability to learn powerful features. Recently, [3] and [4] used CNN-based crowd counting to obtain the estimated density map. Despite the high performance, existing CNN-based counting techniques suffer from algorithmic weaknesses.

First, some CNN-based crowd counting methods used multi-column approach to tackle the scale variation. Most of the previous works tackle the congested scene analysis by using multi-scale architectures [4], [5]. Though they achieved significant performance, these algorithms however, suffer from major shortcomings such as large amount of training time and ineffective branch structure. MCNN [6]

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

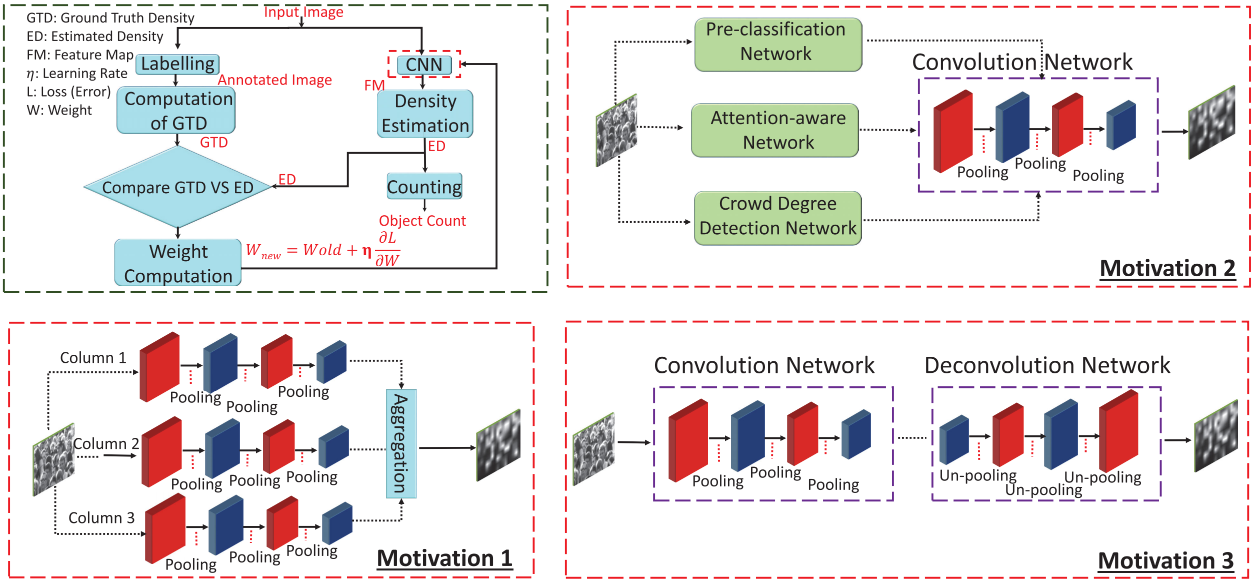

FIGURE 1. Generic form of CNN-based Crowd Counting Technique, Motivations: 1) Parallel model of CNN-based crowd counting, Motivations: 2) Appended networks (pre-classification, attention-aware and crowd degree detection) with CNN-based crowd counting architecture, Motivations: 3) Combination of convolution and deconvolution network foe density estimation.

used multi-column with different kernel sizes to regress the input image and merge different feature maps to obtain the density map. Whereas switch-CNN based crowd counting approach [7] incorporates a switch classifier to forward a patch to specific task-oriented column. While the core of both the methods lies in taking the advantage of characteristics of different receptive fields, the optimal size of column and filter size is still an issue worth addressing. Besides multi-column, the kernel size in each column is fixed which means that column can only handle a specific set of density scenes. Further, three columns in MCNN learn similar type of features irrespective of different filter size [8]. In addition, by taking average of each column's output (density map) further reduces the resolution of final density map (Motivation 1 in Fig. 1). Secondly, most CNN-based crowd counting algorithms focus on the accuracy by neglecting the complexity of overall network [9], [10]. Before CNN-based crowd counting model, different types of complex networks have been appended to increase the accuracy. For example, a pre-classification network classifies the patches of different density level and rest of the CNN-based crowd counting model is trained on specific range of density levels. Attention-aware model also append at the start of CNN-based crowd counting model to obtain the attention region and apply detection or regression task based on the output of earlier model. Moreover, the degree of crowd is estimated to apply a suitable CNN-based crowd counting model. For example, density-aware network that contains multiple sub-networks pre-trained on scenarios with different densities [11]. These supplementary models pre-dating the actual counting architecture enhance the overall complexity, increase the number of parameters, excessively exploit memory making them it impossible to monitor the real time scenes (Motivation 2 in Fig. 1). Lastly, the combination of convolution and deconvolution network in a crowd counting model further enhance, the number of parameters, overall complexity, and memory usage [12], [13]. Instead of using the deconvoluation network to enhance the resolution of density map, a dilated convolution may be effective to obtain the context information by increasing the density map (Motivation 3 in Fig. 1).

Based on these observations, we propose a deeper, wider and more robust approach named Context-aware Scale Aggregation CNN-based Crowd Counting Method (CASA-Crowd). Our model comprises of two types of networks, (i) deep feature extraction network (DFEN) and (ii) scale aggregation module with dilated convolution (SAD). DFEN is responsible for extracting low to complex deep features, whereas, SAD is used to obtain the scale varying features by using GoogLeNet [14] inspired network. In addition, a dilated convolution with varying rates are inserted in inception module [14] to obtain the context information with enhanced resolution of estimated density map. In summary, the main contribution of our research are as follows:

- To the best of our knowledge, it is the first attempt to design a deeper and wider CNN-based crowd counting algorithm. Deeper network is used to capture the simple to complex features, whereas a wider network is responsible to handle scale varying features due to multi-size kernel.
- Combination of DFEN with SAD enhances the ability of network to obtain large scale contextual information, handling the perspective distortion and expanding the spatial sampling location.
- Extensive experiments are conducted on three challenging datasets depicted that our model achieves the state of the art performance.

II. RELATED WORK

With the rapid growth of CNN-based techniques in classification, recognition, and especially segmentation tasks, the CNN-based methods are employed for the purpose of density estimation and crowd counting. To address challenges such as scale-variation due to perspective distortion, non-uniform density distribution, and high variation of density in crowd counting datasets, even at the image level, a number of researchers have contributed to enhance the counting accuracy.

Authors in [15] proposed a multi-column multi-task architecture for drastic scale variation and non-uniform density distribution. Combined density map (Gaussian+human-shaped) is used to obtain the density map. Highly discriminative features are obtained by minimizing per scale loss. However, due to multi-column network, similar type of features are obtained in all columns irrespective of different filter size [8]. Motivated by the success of GANs in an image for image translation problems, authors in [16] employ GANs for crowd counting. The GANs are used for translation of the image and its patches into generated maps. The actual GTD is compared with the generated map to find the best resolution density map (high quality). A novel regularizer adversarial cross scale consistency pursuit network (ACSCP) has been proposed to maintain the parent (whole image) and child (four patches) relationship for reducing the counting loss (previously caused by averaging). By using adversarial loss, the distance between parent density map and concatenated image density map is calculated for minimizing the loss. Authors in [17] proposed a negative correlation learning to enhance the generalizable features. More specifically, the method deeply learns a pool of de-correlated regressors with sound generalization capabilities through managing their intrinsic diversities. Later on, authors in [18] presented residual regression approach to incorporate correlation information among samples. Thus they enable to learn more intrinsic characteristics to enhance the generalization capacity.

Authors in [19] proposed a detection framework for dense crowd counting. They used the multi-column architecture to localize each person in the crowd and spotted heads with bounding box. The enhanced performance shown in terms of localization and counting as well, however, the performance degrades in highly dense and occluded environments. Zhu *et al.* [20] proposed a dual path multi-scale fusion network with attention mechanism. Out of two, one path is responsible to obtain the attention map of crowded regions in an image, whereas the other one is responsible for fusing the scale-varying features and attention map to obtain the high estimated density map. Whereas, VGG-16 network is used as a backbone network to extract the multi-scale features. Tian *et al.* [11] proposed the crowd density-aware network to accurately count the number of people in varying density scenes. It consist of density-aware network, feature enhancement layer, and feature fusion network to improve the accuracy and increases the resolution of estimated density map. To effectively obtain the information of diverse density level, a density-aware network that further contains sub-networks pre-trained on different densities, wherein, feature enhancement layer is used to learn an enhancement rate/weight for each feature maps. Finally, feature fusion network is used to fuse all the feature map. Later on, authors in [21] proposed a dense scale network which consist of pre-trained VGG-16 network as a front-end, where the key component consists of three dense dilated convolution blocks with varying dilation rate to preserve the information from continuously varied scale. However, similar size of filters reduces the capability to obtain large scale context information. Further, the smaller filter size is used throughout the network which results in low performance. As we go deeper into the convolutional pipeline, the background features mix with and dominate the salient features extracted by smaller filter size. In this way, predicting at an earlier layer does not increase the small filter performance as the semantic features are not strong enough for an effective prediction [22].

III. THE PROPOSED APPROACH

The architecture of the proposed approach is shown in Fig. 2. Firstly, CNN-based crowd counting method starts from ground truth density (GTD) estimation as shown in Fig. 1 (top left). Due to varying density, shape, scale and perspective variation in a given dataset, we used different techniques to obtain GTD with varying standard deviation (σ) as shown in Table 1. Secondly, our proposed method employs deep and wide network to obtain different types of features, low to complex, multiple-scale features and contextual information. Thirdly, a dilated convolution convolution is appended in the inception layer benefit from the contextual information while maintaining the quality of density map. In the next sub-section, we explain the computation of ground truth density followed by detailed description of density map estimation.

A. GROUND TRUTH DENSITY ESTIMATION

It is quite difficult to train the CNN model by using the head annotation which is marked as a corresponding dot on each head of an image. Therefore, according to [6], ground truth density is obtained by blurring each of head annotation with Gaussian kernel which is normalized to sum equal to one. In this manner, the sum of density map is equivalent to the total number of objects in an image. Whereas, due to diversified density distribution in different datasets, the accuracy is greatly effected as a result of scale and perspective distorion. Therefore, geometry information is incorporated to reduce the chances of performance degradation. The geometry adaptive kernel (GAK) to calculate the ground truth density can be defined as:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x) \tag{1}$$

$$\sigma_i = \beta \bar{d_i} \tag{2}$$

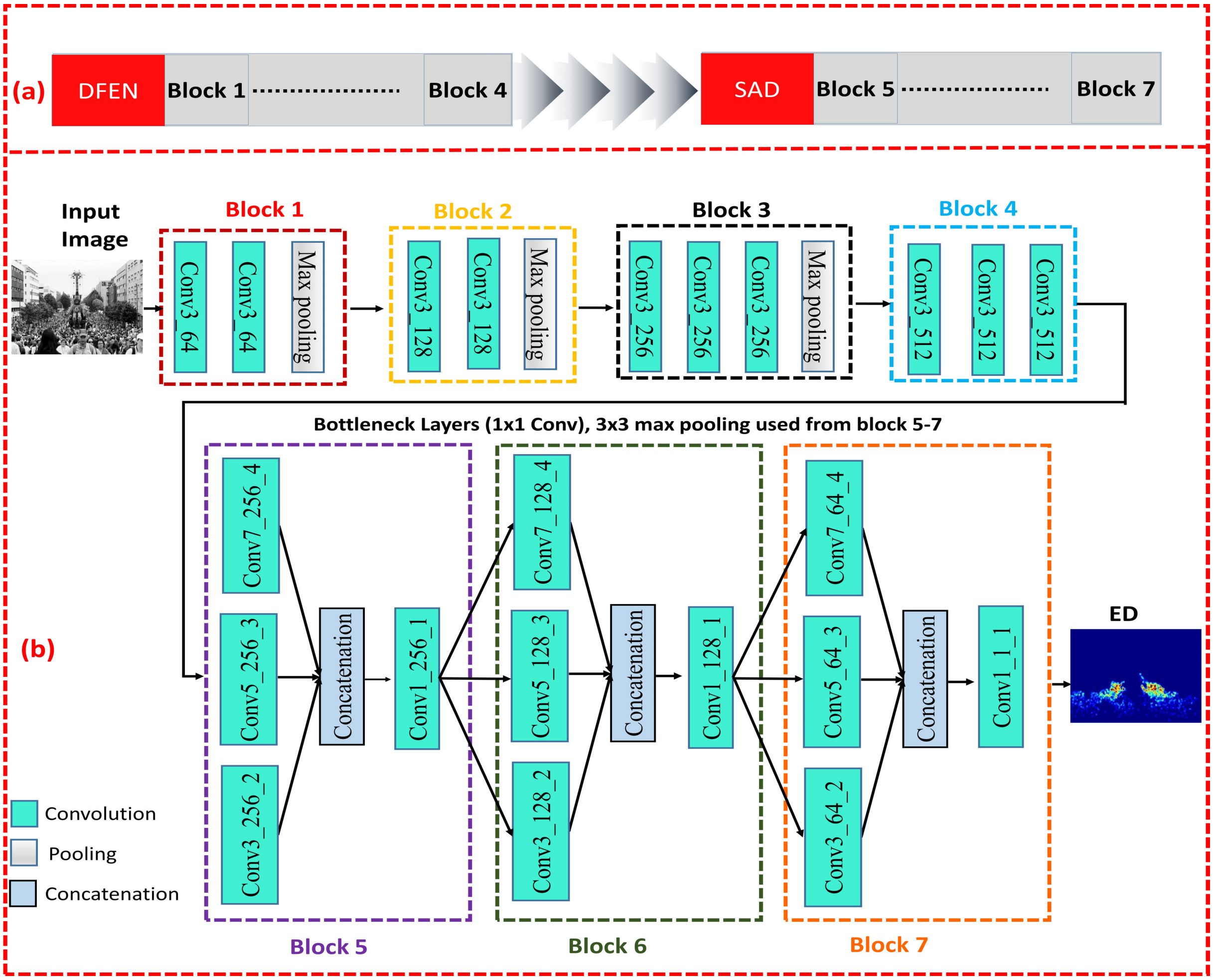

FIGURE 2. (a) The overview of CASA-Crowd, The deep feature extraction network (DFEN), scale aggregation module with dilated convolution (SAD), (b) The whole architecture consist of two parts: one is deep feature extraction network which consist of 4 blocks, the other is Scale Aggregation Module with Dilated Convolution consist of three blocks. The convolutional layers parameters are denoted as "Conv-(kernel size)-(number of filters)-(dilation rate)", max-pooling layers are conducted over a 2 × 2 pixel window with stride of 2.

For each object x_i in the ground truth δ, we use $\bar{d}_i$ as an average distance of k nearest neighbours. To obtain the density map, $\delta(x - x_i)$ convolves with 2-D Gaussian kernel with standard deviation σ_i. Gaussian spread is directly proportional to $\bar{d}_i$ and β which is set to 0.3. We used different values of σ in different datasets as depicted in Table 1.

TABLE 1. The ground truth density generation technique for different datasets.

Dataset	Ground Truth Generating Technique
UCF_CC_50 [23]	GAK
ShanghaiTech Part_A [6]	GAK
ShanghaiTech Part_B [6]	Fixed kernel:σ =15

B. DENSITY MAP ESTIMATION

1) CASA-CROWD:A CONTEXT-AWARE SCALE AGGREGATION CNN-BASED CROWD COUNTING TECHNIQUE

A Context-aware Scale Aggregation CNN-based Crowd Counting method (CASA-Crowd) is depicted in Fig. 2. The proposed network employs two types of network: a network with smaller and same size of filters (block 1- block 4) inspired from VGG-16 named as DFEN. This network is capable of extracting the simple to complex deeper features. When an input is fed to the architecture, it passed through block 1 to block 4 successively. Here, block 5-7 are named as SAD. In each block, an inception module is included with varying dilation rate to obtain the scale and contextual-aware features. With varying receptive field due to different dilation rate, it is very helpful to cope up the perspective variation issues.

a: DEEP FEATURE EXTRACTION NETWORK (DFEN)

One difficulty in crowd counting arises due to variation of density level, background and lack of training data. A large training dataset is required to apply deep learning, however, the largest existing training set contains 1201 images. Due to fewer number of images, it has been conducted by many deep learning models [8], [24] to use pretrained models to avoid overfitting. Most popular backbones such as VGG-16 are trained on ImageNet [25], which is a classification task, whereas crowd counting is a regression task, so the backbone may not directly fit to our model. Meanwhile, Yosinski *et al.* [26] revealed that front-end of the network learns task-independent general features as similar to Gobar filters and color blobs, whereas back-end of the network learns task specified features. Hence, based on these considerations, we choose the first ten convolutional layers of a pretrained VGG-16 as backbone network. Our backbone network is inspired by VGG-16 [27] as shown in Fig. 2 (middle). The network comprises of four blocks, each of which consist of several sequential operations convolution, ReLU and max pooling. Pure convolutional layers are used in backbone to maintain input images resolution. Instead of using the large size filters used by ZFNet [28], combination of two 3 × 3 has the same effective receptive fields as of 5 × 5 [29]. This results in reduced number of parameters with two non-linear units instead of one. We deployed ten layers from VGG-16 in order to extract the low to high level deeper features. Further, the remaining layers are discarded to reduce the computational cost. Blocks 1-2 are responsible to obtain the very low level features like lines, dots, curves etc. Block 3 is used to obtain the complex features like corners, edges, and block 4 is used to obtain the blobs. It is a kind of simple to complex feature learning approach. Further, we choose VGG-16 as the front end of CASA-Crowd due to its strong transfer learning ability. In this way, it is a flexible architecture to concatenate with SAD for density estimation.

b: SCALE AGGREGATION MODULE WITH DILATED CONVOLUTION (SAD)

As shown in Fig. 2 (bottom side, we denote blocks 5 to 7 the scale aggregation module with dilated convolution (SAD). As we know, the pedestrian in the crowd scene usually has contain different sizes due to perspective distortion. Therefore, multi-size filters are necessary to capture features from multiple scales. The SAD is built upon GoogLeNet architecture [14] originally proposed to handle the multiple scales objects simultaneously and provide computational efficiency. The main component of each block consist of an inception module [14] with different size of filter vary from 3×3, 5×5, 7×7 with multiple dilation rate. The output of each branch is concatenated and inserted into next block. The filters are very helpful to convolve the input image on different scales, starting from fine-grained level to the coarse level. Further, filters with small filter size mainly focus to target the small scale, whereas the large size filters are supportive to model the larger targets. From block 5 to block 7, each block with inception module has three branches with varying filter sizes and dilation rate. In this way, SAD module can process feature maps at various scales and aggregate them to the next stage simultaneously, which has been generally proven to be effective in object recognition [14] and image enhancement [30]. Further, varying dilation rate in each branch of each block is helpful to obtain the contextual information. Moreover, dilated convolution is used to enlarge the receptive field. In short, the multi-scale contextual information is aggregated to obtain the state of the art accuracy. As a whole, the SAD deals the multiple-scale variation, perspective distortion, and obtaining the contextual information. In the next sub-section, we will describe the dilated convolution in details.

c: DILATED CONVOLUTION

To address the problem of scale variation, context information aggregation, and quality of density map enhancement, dilated convolution plays a key role. Dilated convolution can be describe as exponential increment of the network's receptive field without an exponential increase in parameters. Further, coverage and resolution of estimated density map is maintained. CNNs with dilated filters have proven to provide competitive performance in image segmentation where multi-scale analysis and contextual information is also critical [31], [32]. Therefore, by incorporating the dilated convolution in SAD, we greatly increase the ability of the network to selectively aggregate multi-scale contextual information, without the need for explicit perspective maps during training and testing.

To perform the perspective-free crowd counting and increasing the spatial sampling location, it is necessary to extract the features from multiple scales. Further, instead of using the larger kernel size which exponentially increase the number of parameters to obtain the multi-scale contextual information, we used the dilated convolution with reduced parameters. Moreover, dilated convolutional layers have been demonstrated in segmentation tasks with significant improvement of accuracy [32] and it is a good alternative of pooling layer. Although pooling layers (e.g., max and average pooling) are widely used for maintaining invariance and controlling overfitting, they also dramatically reduce the spatial resolution meaning the spatial information of feature map is lost.

By observing the above-mentioned benefits of dilated convolution, we incorporate the dilated convolution in SAD inspired from [31]. The traditional 2D convolution can be defined as a real valued function $F : \mathbb{Z}^2 \rightarrow \mathbb{R}$, an input $\Omega_r = [-r, r]^2 \in \mathbb{Z}^2$, and a filter function $k : \Omega_r \rightarrow \mathbb{R}$. The discrete convolution operator $*$ can be defined as:

$$(F * k)(p) = \sum_{s+t=p} F(s)k(t) \tag{3}$$

Let l be a dilation factor and let $*_l$ be defined as:

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{4}$$

We defined $*_l$ as a dilated convolution, whereas a simple convolution $*$ is describe as 1-dilated convolution. The dilated convolution is motivated by the fact of exponentially expansion of receptive field without losing coverage or resolution. Let we define $F_0, F_1, \ldots\ldots, F_{n-1} : \mathbb{Z}^2 \rightarrow \mathbb{R}$ be a discrete functions and let we consider a discrete filters

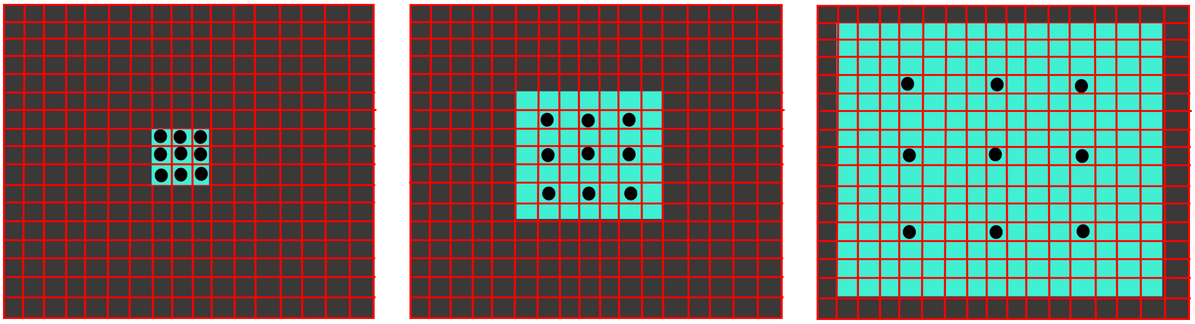

FIGURE 3. The dilated convolution with different receptive fields. Filter size: 3×3, Receptive field: 3×3, 7×7, 15×15.

$k_0, k_1, \ldots\ldots, k_{n-2} : \Omega \rightarrow \mathbb{R}$. By applying the filters with exponential increment of dilation:

$$F_{i+1} = F_i *_{2^i} k_i \quad for\ i = 0, 1, \ldots, n-2. \tag{5}$$

The receptive field of an element p in F_{i+1} as the set of elements on F_0 that modify the value of $F_{i+1}(p)$. Suppose the size of the receptive field of p in F_{i+1} be the number of these elements. The size of the receptive field of each element in $F_{i+1} = (2^{i+2} - 1) \times (2^{i+2} - 1)$. In this way dilation supports exponential increment of the receptive field without effecting resolution. Let us understand systematic dilation through an example shown in Fig. 3. F_1 is produced from F_0 by a dilation convolution and each element in F_1 has a receptive field of 3×3. F_2 is produced from F_1 by a 2-dilated convolution, such that each element in F_2 has receptive field of 7×7. Similarly F_3 is produced from F_2 by a 4 dilated convolution with receptive field of size 15×15. In this way the number of parameters remain constant with exponential increment of receptive field.

IV. EXPERIMENTS

In this section, we describe the whole experiment details starting from network architecture to evaluation of proposed method. Moreover, this section is further sub-divided into three sub-sections: implementation details, architecture ablation and, comparison with state-of-the-art. In addition, we explain the performance comparison of the proposed CASA-Crowd on three well-known datasets.

A. IMPLEMENTATION DETAILS

1) NETWORK CONFIGURATION

The network configuration of CASA-Crowd is shown in Table 2. From block1 to block 4, we used a modified form of VGG-16 network [27], except the fully connected layers (in order to reduce the complexity). Using smaller filter size

TABLE 2. The architecture of CASA-Crowd.

Layer	Channels	Filter	Padding	Max Pooling/Stride	Dilation	CASA-Crowd
Block 1	64	3*3	1	2*2/2	-	Conv3_64 Conv3_64 Max pooling
Block 2	128	3*3	1	2*2/2	-	Conv3_128 Conv3_128 Max pooling
Block 3	256	3*3	1	2*2/2	-	Conv3_256 Conv3_256 Conv3_256 Max pooling
Block 4	512	3*3	1	2*2/2	2/3/4	Conv3_512 Conv3_512 Conv3_512
Block 5	256	3 * 3/ 5 * 5/ 7 * 7	1/2/3	3*3/2	2/3/4	Conv3_256_2 Conv5_256_3 Conv7_256_4 Conv1_256
Block 6	128	3 * 3/ 5 * 5/ 7 * 7	1/2/3	3*3/2	2/3/4	Conv3_128_2 Conv5_128_3 Conv7_128_4 Conv1_128
Block 7	64	3 * 3/ 5 * 5/ 7 * 7	1/2/3	3*3/2	2/3/4	Conv3_64_2 Conv5_64_3 Conv7_64_4 Conv1_64

with more convolutional layers is more efficient instead of applying larger filter size in fewer layers [27]. In this way, block 1 and block 2 have two convolutional layers, block 3 has three convolutional layers followed by a pooling layer with same filter size of 3×3. Whereas, block 4 has only three convolutional layers of similar size. Further, for the extraction of scale-aggregation and contextual information, we append three blocks to obtain perspective varying and context-aware features. From block 5 to block 7, each block has four convolution layers of size (3 × 3/5 × 5/7 × 7), padding (1/2/3) with dilatation rate vary from (2, 3, 4) followed by concatenation of four branches. The similar process is repeated from block 5 to 7 to extract the multi-scale, contextual information with quality of density map.

2) COMPUTATIONAL COMPLEXITY

The redundancy of parameters in deep neural networks is directly related to time and space complexity [33]. We provide a brief discussion on time complexity in terms of the number of parameters. Fig. 4 shows that CASA-Crowd has a comparable computational complexity against [11], [19]–[21]. The reason of this performance is incorporation of SAD modules at the back-end which are based on light weight inception network [34]. Further, DSNet [21] achieves low computational complexity against [11], [19], [20]. This is due to the usage of VGG-16 network only with different dense dilation rate, while CASA-Crowd uses the VGG-16 and inception based modules [14], [27].

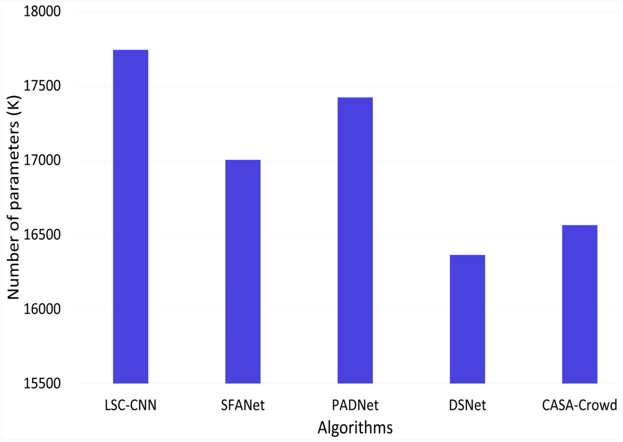

FIGURE 4. **Comparison of computational complexity.**

3) TRAINING DETAILS

The Euclidean distance is used to measure the loss between estimated and ground truth density map. The loss function is given as follows:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|Z(X_i, \Theta) - G_i\|_2^2 \tag{6}$$

where, Θ is a set of parameters. X_i is the input and G_i represents ground truth density. $Z(X_i, \Theta)$ is the output density map for any input X_i, where N is the size of training images. The first 10 layers are fine-tuned from pre-trained VGG-16 architecture [27]. For rest of the network, initial values are taken from a Gaussian initialization standard deviation 0.01. The CASA-Crowd is trained by Stochastic Gradient Descent (SGD) with learning rate 1e-6 and momentum 0.9. We used PyTorch platform [35] with NVIDIA GeForce GTX 1070 with 8GB memory.

4) DATA AUGMENTATION

During the training process, 9 patches are cropped from each image at varying locations, whereas each patch is 1/4 size of original image. The first 4 patches are non-overlapped and it contains four quarter of an image, while the rest of five patches are randomly cropped. To increase the size of training data, we mirror the patches. Data augmentation is not performed for test dataset.

TABLE 3. **Results of the ablation study on ShanghaiTech Part_A dataset.**

	ShanghaiTech Part_A		
Options	GAK	MAE	MSE
DFEN	✓	81.7	121.8
SAD+Without Dilation	✓	84.5	127.3
DFEN+Dilation	✓	68.2	115.2
SAD+Dilation	✓	72.5	121.7
Proposed (CASA-Crowd)	✓	58.6	97.8

B. ARCHITECTURE ABLATION

This subsection is devoted to probing the capability of each component of CASA-Crowd, which specifically indicate the DFEN and SAD module. Due to large variations of crowd density level, we conducted all ablations on ShanghaiTech Part_A dataset. To validate effectiveness of CASA-Crowd, we conduct experiments by adding components incrementally as shown in Table 3. The ablation study consist of five modules added sequentially.

- DFEN: VGG-16 based network.
- SAD+Without Dilation: SAD is an inception-based network without any dilated convolution.
- DFEN+Dilation: VGG-16 based network with dilated convolution as used by authors in [8].
- SAD: Inception-based network with varying dilation convolution.
- DFEN+SAD: Combination of deep and wide network with varying dilation rates.

We evaluate the proposed model by sequentially adding modules. Starting from first DFEN which is based on first 10 layers of VGG-16 network. It achieves MAE of 81.7. Further, we added dilated convolution [8] to analyze the effectiveness of CASA-Crowd. Whereas, we evaluate our next module SAD by eliminating the dilated convolution by achieving a comparable MAE of 84.5. By adding the dilated convolution in SAD, we achieve a significant decrement in MAE. This illustrates that the features with dense, varying scales and large receptive fields caused by incremental dense dilated convolution block are vital and valuable to count crowd accurately and robustly. Finally, by concatenating the DFEN and SAD, we achieve a significant improvement in performance due to scale diversity and enlarge receptive field of features that can handle the issue of large scale variations to perform well on counting the number of people.

C. COMPARISON WITH EXISTING ALGORITHMS

1) EVALUATION METRICS

We evaluate the CASA-Crowd on three well known datasets. For total number of images in test dataset, the loss or count error is obtained by using mean absolute error (MAE) and mean square error (MSE) as given below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{7}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{8}$$

MAE and MSE are evaluation metrics used to estimate the loss. Where N is the number of images in one test sequence, y_i is the estimated count and y' is the corresponding ground truth count.

2) SHANGHAITECH (PART-A)

The ShanghaiTech datasets Part_A [6] is taken from the internet, large scale, largest in terms of number of annotated people, large density as compared to (B), diverse scene and varying densities. Further, Part_A has 482 images (300 images are used for training and 182 used for testing) with 241,677 total number of people, where the minimum, maximum and average head count in an image is 33, 3139 and 501, respectively. We perform comparison with the state-of-the-art algorithms as shown in Table 4. Table 4 shows that the performance in terms of MAE and MSE of CASA-Crowd is comparable to the state-of-the-art methods [11], [19]–[21] when tested on ShanghaiTech (Part-A). The reason is consideration of deeper and wider network to obtain the deep complex features. Further, multi-scale contextual information is obtained to enhance the accuracy. The qualitative results are shown in Fig. 5.

TABLE 4. Estimation errors on ShanghaiTech dataset.

Technique	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN [6]	110.2	173.2	26.4	41.3
C-MTL [10]	101.3	152.4	20.0	31.1
SwitchCNN [7]	90.4	135.0	21.6	33.4
SaCNN [36]	86.8	139.2	16.2	25.8
CP-CNN [24]	73.6	106.4	20.1	30.1
ACSCP [16]	75.7	102.7	17.2	27.4
Deep-NCL [17]	73.5	112.3	18.7	26.0
IG-CNN [37]	72.5	118.2	13.6	21.1
ic-CNN [38]	68.5	116.2	10.7	12.2
CSRNet [8]	68.2	115.0	10.0	16.0
SANet [13]	67.0	104.5	8.4	16.0
LSC-CNN [19]	66.4	117.0	8.1	13.6
DSNet [21]	61.7	102.6	6.7	10.5
SFANet [20]	59.8	99.3	6.9	10.9
PADNet [11]	59.2	98.1	8.1	12.2
CASA-Crowd (Proposed)	58.6	97.8	7.8	12.5

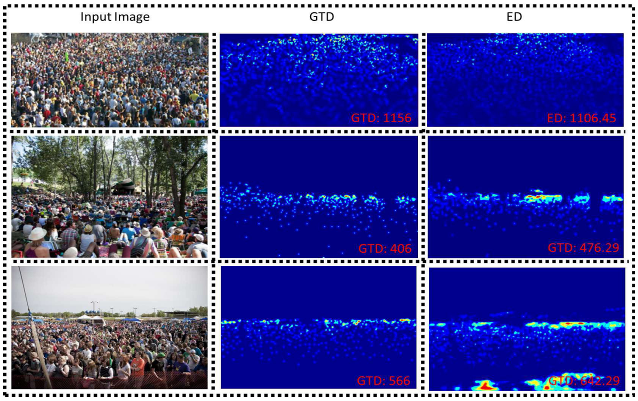

FIGURE 5. Visualization of ShanghaiTech Dataset (Part_A), ground truth density, estimated density.

3) SHANGHAITECH (PART-B)

The ShanghaiTech datasets Part_B [6] contains total 716 images (400 images are used for training and 316 are used for testing) with 88,488 total number of people with minimum, maximum and average head count equal to 9, 578 and 123, respectively. Moreover, it is collected from Shanghai with varying scale and perspective. Non-uniform density level in many images make it tilt towards the low density level. We compare CASA-Crowd with existing algorithms [11], [19]–[21] on ShanghaiTech datasets Part_B. The performance of CASA-Crowd in terms of MAE and MSE is shown in Table 4. MAE and MSE of CASA-Crowd is comparable to the state-of-the-art methods [20], [21] on Part_B. The reason for such error enhancement is due to overestimation of crowd density in sparse density areas. Further, the qualitative results are shown in Fig. 6.

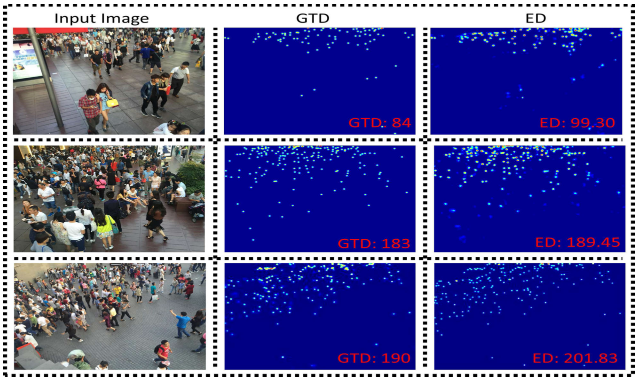

FIGURE 6. Visualization of ShanghaiTech Dataset (Part_B), ground truth density, estimated density.

4) UCF_CC_50

UCF_CC_50 [23] is collected from various places like concerts, marathon, diverse scene with wide range of densities,

TABLE 5. Estimation errors on UCF_CC_50 dataset.

Technique	MAE	MSE
MCNN [6]	377.6	509.1
C-MTL [10]	322.8	341.4
SwitchCNN [7]	318.1	439.2
SaCNN [36]	314.9	424.8
CP-CNN [24]	295.8	320.9
ACSCP [16]	291.0	404.6
Deep-NCL [17]	288.4	404.7
IG-CNN [37]	291.4	349.4
ic-CNN [38]	260.9	365.5
CSRNet [8]	266.1	397.5
SANet [13]	258.4	334.9
LSC-CNN [19]	225.6	302.7
SFANet [20]	219.6	316.2
PADNet [11]	185.8	278.3
DSNet [21]	183.3	240.6
CASA-Crowd (Proposed)	182.7	285.4

challenging dataset as compared to ShanghaiTech. The dataset consists of 50 images, where the number of people per image start from 94 to 4543 with an average number of 1279. Furthermore, 5-fold cross validation is performed according to standard settings in [23]. Table 5 depicts that the MAE of [11], [19]–[21] is high (low accuracy) as compared to CASA-Crowd. Additionally, CASA-Crowd takes advantage of larger receptive field to obtain the context information, whereas perspective distortion is improved by extracting the low to complex features by using DFEN and SADM. The qualitative results are shown in Fig. 7.

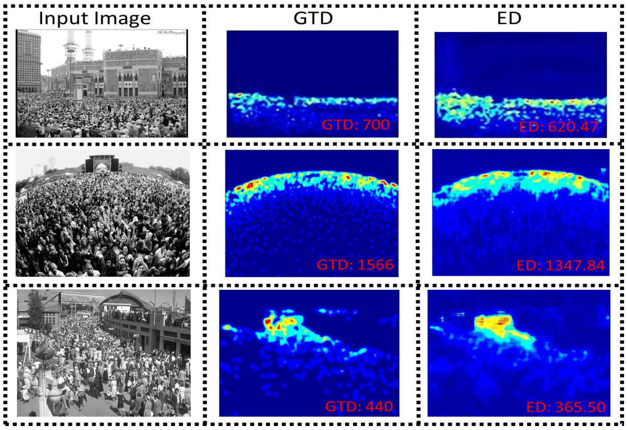

FIGURE 7. Visualization of UCF_CC_50 Dataset, ground truth density, estimated density.

V. CONCLUSION AND FUTURE WORK

In this work, we proposed a novel architecture called a context-aware scale aggregation CNN-based crowd counting method (CASA-Crowd) that is trained in an end-to-end manner. Due to strong feature extraction property of deep neural network, we used deeper and wider networks to extract the deep and scale varying features. Furthermore, a dilated convolution approach is included in inception module to obtained the context-information. The performance of CASA-Crowd is comparable to the state-the-art methods due to varying receptive field and strong ability of handling the perspective varying issues. Moreover, the quality of density map is enhanced due to expanded spatial sampling. In this way, our proposed approach is capable of learning the low to complex, deeper and scale-aware features with enhanced density map. In future, we intend to incorporate the segmentation with crowd counting to further observe different classes within any scene to improve the accuracy of overall crowd counting.

REFERENCES

[1] Z. Cheng, L. Qin, Q. Huang, S. Yan, and Q. Tian, "Recognizing human group action by layered model with multiple cues," *Neurocomputing*, vol. 136, pp. 124–135, Jul. 2014.

[2] A. Marana, L. D. F. Costa, R. Lotufo, and S. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. IEEE Int. Symp. Comput. Graph., Image Process., Vis. (SIBGRAPI)*, Oct. 1998, pp. 354–361.

[3] Y. Zhang, C. Zhou, F. Chang, and A. C. Kot, "Multi-resolution attention convolutional neural network for crowd counting," *Neurocomputing*, vol. 329, pp. 144–152, Feb. 2019.

[4] Y. Bharti, R. Saharan, and A. Saxena, "Counting the number of people in crowd as a part of automatic crowd monitoring: A combined approach," in *Information and Communication Technology for Intelligent Systems*. Singapore: Springer, 2019, pp. 545–552.

[5] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 615–629.

[6] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 589–597.

[7] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jul. 2017, pp. 4031–4039.

[8] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[9] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," 2018, *arXiv:1807.00601*. [Online]. Available: https://arxiv.org/abs/1807.00601

[10] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug./Sep. 2017, pp. 1–6.

[11] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "PaDNet: Pan-density crowd counting," 2018, *arXiv:1811.02805*. [Online]. Available: https://arxiv.org/abs/1811.02805

[12] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, "Crowd counting and density estimation by trellis encoder-decoder network," 2019, *arXiv:1903.00853*. [Online]. Available: https://arxiv.org/abs/1903.00853

[13] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[15] B. Yang, J. Cao, N. Wang, Y. Zhang, and L. Zou, "Counting challenging crowds robustly using a multi-column multi-task convolutional neural network," *Signal Process., Image Commun.*, vol. 64, pp. 118–129, May 2018.

[16] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5245–5254.

[17] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5382–5390.

[18] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu, "Residual regression with semantic prior for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4036–4045.

[19] D. B. Sam, S. V. Peri, A. Kamath, and R. V. Babu, "Locate, size and count: Accurately resolving people in dense crowds via detection," 2019, *arXiv:1906.07538*. [Online]. Available: https://arxiv.org/abs/1906.07538

[20] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual path multi-scale fusion networks with attention for crowd counting," 2019, *arXiv:1902.01115*. [Online]. Available: https://arxiv.org/abs/1902.01115

[21] F. Dai, H. Liu, Y. Ma, J. Cao, Q. Zhao, and Y. Zhang, "Dense scale network for crowd counting," 2019, *arXiv:1906.09707*. [Online]. Available: https://arxiv.org/abs/1906.09707

[22] B. A. Mudassar and S. Mukhopadhyay, "Rethinking convolutional feature extraction for small object detection," in *Proc. BMVC*, 2019.

[23] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[24] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1861–1870.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[28] P. Hao, J.-H. Zhai, and S.-F. Zhang, "A simple and effective method for image classification," in *Proc. IEEE Int. Conf. Mach. Learn. (ICMLC)*, vol. 1, Jul. 2017, pp. 230–235.

[29] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, and J. Cai, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[30] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[32] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[33] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S.-F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2857–2865.

[34] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop Autodiff Submission*, 2017.

[36] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1113–1121.

[37] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3618–3626.

[38] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 270–285.

NAVEED ILYAS received the bachelor's degree in computer engineering and the master's degree in electrical engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2011 and 2015, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), South Korea. He has authored more than ten research articles in international journals and conferences. He is currently involved in scientific and instructive research in the fields of machine learning, deep learning, computer vision, and crowd analysis.

ASHFAQ AHMAD (S'15) received the B.S. degree in electrical (telecommunication) engineering and the M.S. degree in electrical engineering with a specialization in wireless sensor networks and energy management in smart grid from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree in electrical engineering from The University of Newcastle, Australia. He has (co)authored more than 40 research publications in international technical journals and peer reviewed conferences. His research interests include modeling, analysis, and control of hierarchical, hybrid, and stochastic systems, with applications in energy/power systems and wireless networks. He has served and is actively serving as a Guest Editor, an Invited Reviewer, and a TPC Member of prestigious international journals and conferences/workshops.

KISEON KIM (M'84–SM'98) received the B.Eng. and M.Eng. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1978 and 1980, respectively, and the Ph.D. degree in electrical engineering systems from the University of Southern California, Los Angeles, CA, USA, in 1987. From 1988 to 1991, he was with Schlumberger, Houston, TX, USA. From 1991 to 1994, he was with the Superconducting Super Collider Laboratory, TX. In 1994, he joined the Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently a Professor. His current research interests include biomedical applications design, wideband digital communications system design, sensor network design, and analysis and implementation both at the physical layer and at the resource management layer. He is a member of the National Academy of Engineering of Korea, a Fellow of the IET, and the Senior Editor of the IEEE Sensors Journal.