IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.Doi Number

Crowd Density Estimation by Using Attention Based Capsule Network and Multi-Column CNN

MERVE AYYUCE KIZRAK[1, 2], and BULENT BOLAT[2]

[1] Department of Artificial Intelligence Engineering, Faculty of Engineering and Natural Sciences, Bahçeşehir University, 34353 Istanbul, Turkey
[2] Department of Electronics and Communication Engineering, Faculty of Electrical and Electronics Engineering, Yıldız Technical University, 34220 Istanbul, Turkey

Corresponding author: Merve Ayyuce Kizrak (e-mail: merveayyuce.kizrak@eng.bau.edu.tr).

ABSTRACT We propose a strategy that focuses on estimating the number of people in a crowd, one of the aims of crowd analysis, using static images or video images. While manual feature extraction was not performed with pixel and regression-based methods in the first studies on crowd analysis, recent studies use Convolutional Neural Networks (CNN) based models. However, it is still difficult to extract spatial information such as position, orientation, posture, and angular value for crowd estimation from a density map. This study uses capsule networks and routing by agreement algorithm as an attention module. Our proposed approach consists of both CNN and capsule network-based attention modules in a two-column deep neural network architecture. We evaluate our proposed approach compared with other state-of-the-art methods using three well-known datasets: UCF-QNRF, UCF_CC_50, UCSD, ShangaiTech Part A, and WorldExpo'10 .

INDEX TERMS Capsule attention, crowd counting, density map, multi-column CNN.

I. INTRODUCTION

Population growth and rapid urbanization gather people together and require planning in order to prevent crowd congestion. While the management and surveillance of provocative events in city squares in the past years have been drawing attention, counting people at points such as queues and shopping areas and monitoring their contact situations have become critical for world health due to the Covid-19 pandemic in the last year.

Crowd management covers many topics such as flow analysis, urban planning, public safety management, disaster management, and defense to prevent congestion in events and open spaces. Crowd analysis is a tool for crowd management. It examines the number, distribution, and behavior of people in various scenes using images and videos. Analysis of the numbers and spatial distributions of people is basically grouped into recognition, tracking, and counting.

Homographic, pixel and color-based approaches project local image features from each sensor onto a common plane to provide general object detection. Until the effectiveness of CNN-based approaches was proven in the literature, homographic, pixel, and color-based approaches were used for people recognition, tracking, and crowd counting. Such approaches are particularly vulnerable to variables such as light and size [1]-[6]. A deep neural network is used to enable perception by utilizing end-to-end multiview
information. However, in the people counting problem, there are frontal, profile, and overlapping scenes, and therefore, most studies focus on solving these problems. For this, some studies focus on mean area variational inference, probability occupancy mapping (POM), and estimating the posterior probability distribution of people at the scene. One of the latest studies being discussed uses Conditional Random Fields (CRFs) for multi-view pedestrian detection [7] - [13]. However, crowd analysis remains a research area that requires further study. Convolutional Neural Network (CNN) based studies are frequently recommended because of the path and pattern learning ability in the field of computer vision. Every day, new research is undertaken to determine the number of people, their behaviors, and their sizes in images, and to count people in very dense crowds.

In the light of these observations, our study proposes an attention-based model which uses the ability of the two-column CNNs to learn useful features and also uses the spatial information acquisition feature of the Capsule Networks (CapsNet). The effect of CapsNet, which has not been widely investigated for crowd analysis as an attention module, is examined on various datasets and shows that the results are comparable with current studies in the literature.

The main contributions of our research can be summarized as follows:

- Although object recognition-based CNN approaches to crowd counting constitute an important approach that has far exceeded the success of pixel and color-based approaches in the past, they have unresolved problems in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

taking on spatial information that varies according to scale.

- Predictions for spatial information cannot be solved by the state of the art CNNs with a CapsNet-based approach; using the routing by agreement algorithm in case of position, orientation, posture, and angular value change, the aim is to learn the spatial features representing the object. In this study, estimation is made on the spatial values of the crowded group. The unique orientations, postures, and angular values of individuals and groups in crowded images are important in predicting crowd density estimation. Using these spatial features, crowd behavior analysis may be the subject of future studies.
- Experiments on five challenging datasets demonstrate that the method we propose achieves the best performance among recent studies except for Mean Absolute Error on the UCF-QNRF dataset.
- The effects of kernel selection and model architecture of the proposed method with ablation studies are also examined through experiments. Through model architecture ablation studies, the contribution of using CapsNet and CNN attention mechanisms separately and together has been shown.

The remainder of the paper is organized as follows: after the related work in Section II, we cover the details of our proposed method in Section III. Section IV introduces our datasets, evaluation metrics, experimental results, and ablation studies. We conclude with a short discussion in Section V.

II. RELATED WORK

Image analysis and crowd counting approaches are a field of study that draws the attention of researchers and finds a counterpart in practical life. In this paper, we research CNN and attention-based crowd analysis and counting approaches.

A. CNN-BASED APPROACHES

Depending on the regression between image features and crowd size, regression-based approaches can be used to estimate the number of people in crowds. The regression approach is used to extract density maps from the image with CNN-based approaches. Zhang *et al.* [14] used the CNN regression model to estimate the number of people in a single image in two configurations. Zhang *et al.* [15] proposed a multi-column CNN model to prevent distortions arising from a perspective using the MCNN. Sam *et al.* [16] selected Switch-CNN, the crowd counting architecture with the highest performance, and made evaluations over a multi-column CNN architecture. Sindagi *et al.* [17] used Pyramid from Context with the CNN model to estimate crowd density numbers with a high degree of accuracy by using global and local contextual information. Shi *et al.* [18] focused on learning a negative correlation to develop generalizable features. With this method, learning consisting of unrelated regressors with robust generalization abilities was achieved by managing their internal diversity. Zhu *et al.* [19] used different regression networks to calculate crowd density.

Although regression-based approaches are successful in density estimation, they do not perform well enough in low-density situations and when localization is required. CNN-based approaches can successfully perform tasks such as classification, recognition, and segmentation in many areas. They are also used in crowd analysis studies such as density estimation, crowd counting, localization, tracking, and surveillance. Different CNN approaches are also included in current studies to overcome difficulties such as perspective distortions and non-uniform density changes.

The first handling of the problem of counting people from crowd images and videos using CNN was realized by studies [20] and [21]. While the CNN regression model was used for the pedestrian counting task in [20], a classification process at five levels was carried out on the density image in [21]. [22], which adopts a patch-based approach over a single image for the person counting task, proposes an end-to-end estimation method. The CNN approach, in which the most suitable regression can be selected automatically, is among the current studies in an inspiring method for combining the people counting task with other tasks [23]. Shang *et al.* [24] emphasize the solution to the problem of decreasing accuracy in changing scenes with the cross-scene counting model. Successful results were obtained by capturing semantic information from the image in the CrowdNet study, which suggested a combination of shallow and deep CNN models [25]. Similarly, a cascaded CNN model has been proposed that can predict density mapping and classify people counted at different intensity levels simultaneously [26]. Mundhenk *et al.* [27] use density maps produced by the people counting model to carry out counting, tracking, and perception tasks together. A method that combines counting people to calculate the speed of people passing by in a higher-level cognitive task is proposed in [28]. Sindagi *et al.* [29] and [30] propose approaches where both people counting and density estimation can be classified simultaneously with the ResNet-based model. In a context-aware scale aggregation CNN-based crowd counting technique (CASA-Crowd) [31], the focus was on obtaining features that change with depth, changing scale, and perspective. Also, an extended convolution with varying filter sizes was used to obtain contextual information. On the other hand, due to different dilation ratios, a variation in receptive field size causing perspective distortion was overcome.

Zhang *et al.* [32] used an approach called CSRNet and aimed to expand the feature area and create quality density maps by making modifications in pooling operations. Shi *et al.* [33] proposed the perspective-aware CNN model to reduce the most common perspective problem in crowd prediction. Wan *et al.* [34] employed a residual regression approach using correlation information between samples, thus aiming to learn more intrinsic characteristics in order to increase the generalization capacity. Dai *et al.* [35] employed a model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

which consists of three convolution blocks to withstand varying expansion rates and selected smaller filter sizes than in other studies to capture contextual information. The pre-trained VGG-16 network is also used as a base model in this study [35]. Jiang *et al.* [36] proposed a Multi-Level Convolutional Neural Network (MLCNN) architecture which first learns multilevel density maps adaptively and then combines them to estimate the number of people in crowds. Oñoro-Rubio *et al.* [37] proposed a model based on the MCNN [15] model, which works by extracting features in different resolutions and without using perspective information on crowd distribution and crowd number to overcome perspective distortions. On the other hand, MCNN, a multi-branched CNN approach, was recommended in [38], where a three-dimensional filtering approach was used to learn the features of the image, and thereby aimed to analyze different levels of information. Yang *et al.* [39] proposed a multi-column CNN architecture for variable density maps and suggested that they achieved more successful results in this way. Tian *et al.* [40] suggested a method using density mapping for counting people in scenes with varying densities. In this study using a feature fusion network, information on different intensity levels was obtained more effectively. Gao *et al.* [41] proposed a Perspective crowd counting CNN (PCCNet) and crowd counting method to reduce errors arising from high similarity and perspective changes in appearance. Sam *et al.* [42] used a multi-column CNN architecture to position each person in the crowd for dense crowd counting. This study also used the bounding box method to locate people's heads in the crowd. The performance ranged from successful in medium crowd images to not robust in very dense images. Guo *et al.* [43] used scale collection modules for high-resolution density maps to solve the scale diversity problem. Shen *et al.* [44], based on the success of GAN models in image distortion, used adversarial cross-scale consistency pursuit (ACSCP) with four sub-patches on a high-quality density map for crowd counting. Using the contrast loss, the distance between the main density map and the merged image density map was calculated to minimize the loss. Liu *et al.* [45] proposed a combined method that includes both regression and perceptual counting and adaptively decides on the appropriate counting mode for different image positions. Object detectors in detection-based methods can determine the position of each person. Thus, it enables crowd counting and localization [46], [47], [48]. The model proposed in [49] is similar to the head detector training model in [32]. Instead of using the general object sample into the network as suggested in [50], it offers scale-sensitive samples using a scale map. Scale maps can estimate object scales, and accordingly, this is a more effective approach to direct suggestions than is making comprehensive searches in all scales.

B. ATTENTION-BASED APPROACHES

Hu *et al.* [50] suggested an attention mechanism called SENet to focus on valuable features in the image. More successful results were obtained by combining channel and spatial attention mechanisms in [51] and [52]. Although attention models were first used in the field of natural language processing such as machine translation, they have been achieving significant successes for a while in studies such as image-based object detection, classification, segmentation, and face recognition [53], [54], [55], [56]. By combining the advantages of [59] and [50] and [57], the method proposes a global context module, and thus image classification is also a significant success. Li *et al.* [58] showed that an attention mechanism that can dynamically select target sizes of neurons can provide successful results on images. [60] ADCrowdNet, which consists of two CNN networks, first predicts crowded areas of the image, and this attention model then generates high-quality density maps. A feature fusion attention network (FFANet) is recommended for crowd counting [61]. FFANet is implemented in conjunction with the VGG16 network, and features extracted from crowd images are combined. Knowledge development operations on multi-level features are carried out by Feature Fusion Attention Module (FFAM), which is now further enhanced by Block (RB). These properties are processed by the Compression Module (CM) to create a density map.

III. THE PROPOSED METHOD

In crowd analysis, we propose an innovative approach to create a robust density map. Using CNNs and CapsNet together with an attention mechanism, a more efficient density map for crowd estimation is obtained. CNNs' success in crowd analysis is quite accurate. However, there are limited resources available in the literature regarding the application of CapsNet in this field, due especially to the complexity of the process and the novelty of this approach. Nguyen *et al.* [62] used CapsNet in addition to CNN while determining types of attacks in static and video images. Algamdi *et al.* [63] used a CapsNet architecture to recognize human action using frames from videos without explicit motion information.

In the proposed approach the pre-trained VGG-16 model is used as the base model. This model is divided into two columns after the 4th block of VGG-16; thus, a two-level feature map is obtained. In the second column, there are two different modules, the convolution and capsule attention mechanism modules. CapsNet learns to span the space of variation in scenes. These variations include density, sparseness, and direction of the crowd. Viewpoint invariant knowledge results from CapsNet transformation matrices that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

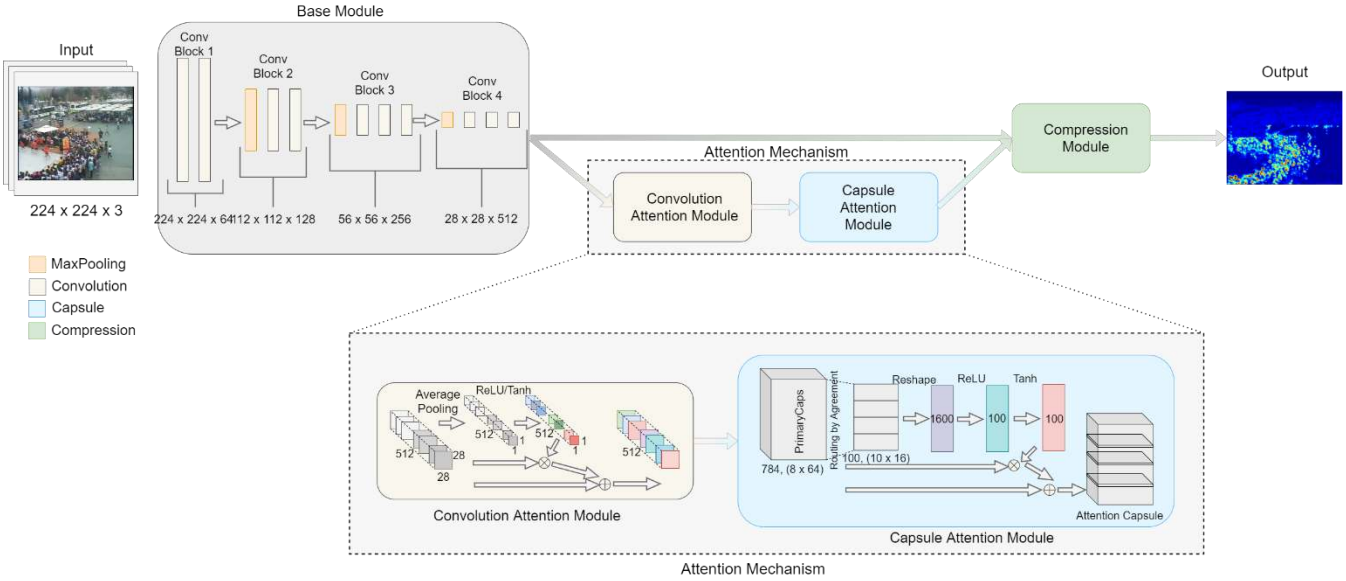

FIGURE 1. Overall architecture representation of the proposed model. The resized images are used in the base module input. Then the model is divided into two columns and the convolutional and capsule attention modules are cascaded in one column. The last module consists of compressing two columns.

have learned to encode the inherent spatial relationship between a part and a whole, and this knowledge automatically generalizes to novel viewpoints. Using CapsNet, which is robust to affine transformations, more stable results were expected to be obtained for crowd density estimation. While the architecture proposed for this purpose benefits from CNN's success and speed in object recognition, CapsNet's dynamic routing algorithm is used as a parallel attention mechanism that allows each capsule of one level to participate in some active capsules of the following level and ignore the others [68]. The routing-by-agreement algorithm allows a prior view of the shape of a crowd to be used to support the segmentation of the individuals in the scene, and it eliminates the need to make higher-level decisions in the pixel domain. This should allow the proposed model to recognize multiple people in the crowd image even if people mostly overlap. By combining CNNs and the CapsNet attention approach, we avoid the inefficiency of the CNN model. This inefficiency is expected to increase exponentially due to overlap. In the last stage, a density map is created using the feature maps obtained by compressing the base model and both attention modules. It is suggested that the CapsNet attention mechanism, which can keep spatial and temporal information, should be applied together with CNNs and used for crowd analysis by going beyond its original purposes.

A. BASE MODULE

The first 4 convolutions comprise a pre-trained VGG-16 network in which blocks have been used and fully connected layers have been removed [64]. A BN layer has been added behind all convolutional layers in the VGG-16 network. The base model consists of 10 convolutional layers. Except for the first convolution block, the max-pooling operation has been used in the others. The input images have been resized to 224×224×3. The simple feature map of 28×28×512 size obtained as a result of convolution processes continues to be processed in two columns. It is used directly in the compression module of a column. The other column is used as the input of the convolution attention module. The proposed model architecture is shown in Figure 1.

B. CAPSULE NETWORK

The basic concept is presented in experiments by Hubel and Wiesel [65], and modeled by Fukushima [66], and Lecun *et al.* CNNs, the first successful application of which was developed by [67], are frequently preferred in computer vision applications due to their accurate results.

Feature maps of the image are obtained at the outputs of the convolution layers in the CNNs. The dimension reduction performed by average-pooling also causes information loss. Besides, the overlapping of the objects causes the CNNs to have difficulties in object recognition, classification, and segmentation [68]. CapsNet and dynamic routing algorithm [68], [69] have been proposed as a solution to the problems for which CNNs are insufficient. It shows the decoding structure with a layer called DigitCaps. DigitCaps consists of two full connectivity layers controlled by ReLU and tanh. The Euclidean distance between the images used in training and the output of the sigmoid layer is minimized. DigitCaps has a strategy based on using the correct label as a reconstruction target in training.

CapsNet'in s_j input is shown in equation (1):

$$s_j = \sum_i c_{ij} \hat{p}_{j|i} \tag{1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

The total value of the in a capsule s_j input is a weighted sum of all prediction vectors $\hat{p}_{j|i}$from the capsule of the previous layer. Here $\hat{p}_i$ is the output of the previous layer multiplied by the weight matrix W_{ij} and is calculated as in equation (2).

$$\hat{p}_{j|i} = W_{ij}.\hat{p}_i \tag{2}$$

Where v_j, jth is the output vector of the capsule, s_j, jth is the input vector of the capsule, $\|s_j\|$ is the module of the vector length s_j. The length of the output vector of the capsule layer represents the probability that the entity represented by the capsule is present in the input. For this reason, a nonlinear activation function is used to compress the short vector to a length close to 0 and the long vector to a length less than 1 and close to 1. It is scaled to 0.5 in this study instead of 1 for the squash function. Therefore, the squashing equation is changed as in (3). With this approach, it has been shown that the accuracy of CapsNet increases by between 0.5-1.59% [70].

$$v_j = Squash(s_j) = \frac{\|s_j\|^2}{1+\|s_j\|^2} \Longrightarrow \frac{\|s_j\|^2}{0.5+\|s_j\|^2} \tag{3}$$

For the PrimaryCaps layer, the initial value of b_{ij} selected as 0 should be calculated as in equation (4). In PrimaryCaps layer output c_{ij} the coupling coefficient is defined as equation (5) and an iterative dynamic routing process is performed.

$$b_{ij} + \hat{p}_{j|i}.v_j \Longrightarrow b_{ij} \tag{4}$$

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})} \tag{5}$$

Although difficult to solve by classical CNNs with CapsNet, the thickness, scale, shift, etc. representing the object can be used to successfully recognize the object through capsules consisting of a group of neurons, even when the position, orientation, posture, and angular value change. It has been suggested that the characteristics be learned by routing-by-agreement [68].

C. ATTENTION MODULES

CNNs, recurrent neural networks, and attention models are successful applications for image and natural language processing. In this study, the attention module is used in both the CNNs and the CapsNet. Thus, the desire is to benefit from the robustness and the spatial information of the CapsNet used for this purpose for the first time in crowd analysis.

1) CONVOLUTIONAL ATTENTION MODULE

The input of the convolution attention module is taken as 28×28×512 from the 10th convolution layer of VGG-16, which was previously used as the base model. Feature map convolution is obtained by averaging a fully connected network that sums ReLU and plane information into point information and then uses the ReLU and tanh activation functions respectively. In the last step, the attention convolution is multiplied. The detail of the convolution attention module is shown in Figure 2. Here, the convolution input dimensioning equation p_{sc} is represented by the length M, the width N, and the number of channels Q. The adjustment value used after the average pooling is calculated as in p_a equation (6). Its size is obtained as 1×1×Q. The aim here is to improve the learning ability of the model by using nonlinear functions. Two fully connected layers are applied to p_a, one with the activation function ReLU (p_1) and the other with tanh (p_2). The weight matrices corresponding to the fully connected layers are shown as W_1 and W_2, while the offset is b_1 and b_2. The size of p_2 obtained as a result of these two fully connected layers is 1×1×Q and is shown in equations (7) and (8). The p_3 equation (9) is obtained from the product of p_2 obtained from the input and p_{sc}. In the next step, the convolution attention $p_{conv-att}$ equation (10) is calculated by summing p_3 and p_2.

$$p_{aq_k} = \frac{1}{|MN|}\sum_{i=0}^{M}\sum_{j=0}^{N} q_{kij} \tag{6}$$

$$p_1 = ReLU(W_1 p_a + b_1) \tag{7}$$

$$p_2 = tanh(W_2 p_1 + b_2) \tag{8}$$

$$p_3 = p_{sc} * p_2 \tag{9}$$

$$p_{conv-att} = p_{sc} + p_3 \tag{10}$$

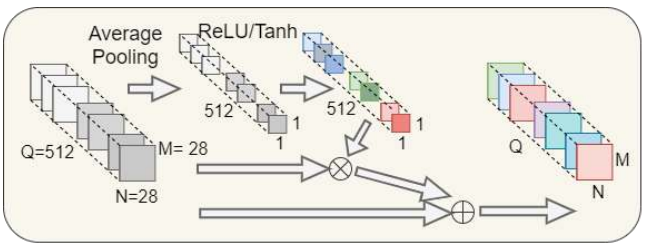

FIGURE 2. Convolutional attention module.

2) CAPSULE ATTENTION MODULE

In order to create the Capsule attention map, it is necessary to calculate high-level DigitCaps using PrimaryCaps. The major disadvantage of CapsNet is the need for processing power. To avoid this problem, one-dimensional convolution and linear activation functions are applied in fully connected layers [70]. Thus, the PrimaryCaps size is obtained as 100×(10×16). PrimaryCaps are reshaped and vectorized. Thus, using the ReLU and tanh activation functions, it is simulated to a fully connected neural network and calculated as in equations (11) and (12).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

TABLE 1. Descriptions of the crowd datasets [71].

Datasets	Description	#of Images	Resolution	Min	Ave	Max	Overall Count	Accessibility
UCF-QNRF [72]	Crowd Counting and Localization	716	400×300 9000×6000	9	123	578	88488	Yes
MALL [73]	People counting	2000	320 × 240	13	-	53	62325	Yes
UCF_CC_50 [74]	Density Estimation	50	Variable	94	1279	4543	63974	Yes
UCSD [75]	People counting	2000	238 × 158	11	25	46	49885	Yes
Shanghai Tech Part A, Part B [76]	Crowd Counting	482	768×1024	33	501	3139	241677	Yes
WorldExpo'10 [77]	Cross Scene Crowd Counting	3980	576 × 720	1	50	253	199923	Yes
CUHK [78]	Crowd Behavior	1535	Variable	49	815	12865	-	Yes

Here p_{s1} and p_{s2} are the results of two fully connected layers, W_3 and W_4 are the weight matrices of these layers, and b_3 and b_4 offset values. For the last step of this module, p_{s3} is calculated from the products of p_s and p_{s2} as in equation (13). Here, by summing p_s and p_{s3}, the M×N×Q dimensional capsule attention map called $p_{caps-att}$ is obtained as shown in equation (14). The detail of the capsule attention module is shown in Figure 3.

$$p_{s1} = ReLU(W_3 p_{sr} + b_3) \quad (11)$$

$$p_{s2} = tanh(W_4 p_{s1} + b_4) \quad (12)$$

$$p_{s3} = p_s * p_{s2} \quad (13)$$

$$p_{caps-att} = p_s + p_{s3} \quad (14)$$

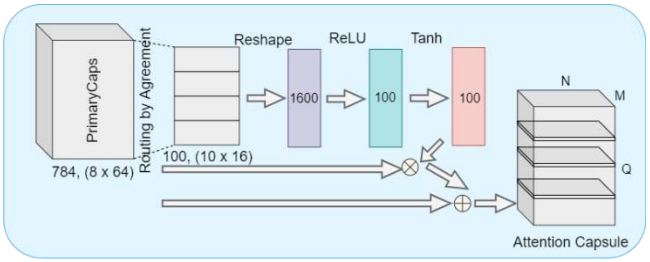

FIGURE 3. Capsule attention module.

3) COMPRESSION MODULE

After obtaining the spatial and temporal information among multi-level features with two different attention modules, the two attention maps as CM are compressed into a single channel crowd density map, as shown in equation (15).

$$CM = [p_{conv-att}, p_{caps-att}] \quad (15)$$

IV. EXPERIMENT and ABLATION STUDIES

In this study, the datasets used for crowd analysis in the literature are summarized and the model we propose with three different datasets is compared with the recent studies in the literature. Evaluation metrics for performance analysis and ablation studies for the proposed model are also included.

A. DATASETS

Video and static images collected for various purposes and from different areas for crowd analysis are shared as open datasets. These images are used to test the performance of implementations for tasks such as estimating the number of people, density estimation, localization, and behavior analysis, and to compare them with other studies in the literature. Datasets used for crowd analysis studies are shown in Table I.

The UCF-QNRF dataset is a more recent dataset prepared for crowd counting and localization. It is challenging due to the high-density variation and resolution change [72]. The MALL dataset consists of images collected from a security camera in a shopping mall. Generally, it is used indoors for counting people [73]. UCF_CC_50 is a challenging dataset used for density mapping, containing information on real organization moments in stadiums, squares, and concert venues, and the gathering and dispersal times of crowds of different densities [74]. The UCSD dataset is collected from images taken from streets that are public environments for the people counting problem [75]. The ShanghaiTech dataset consists of two parts, named A and B. Part A consists of images taken randomly from the internet, and Part B consists of images collected from a metropolitan street in Shanghai [76]. WorldExpo'10 is a dataset prepared for inter-scene crowd counting but its use for accuracy assessment is insufficient due to its low density [77]. CUHK is a published dataset for behavioral analysis obtained from airports, shopping malls, parks, and streets [78].

The UCF-QNRF, UCF_CC_50, UCSD, ShangaiTech Part A, and WorldExpo'10 datasets were used for this study. These datasets were resized and adapted to base model input in order not to be adversely affected by inputs with different resolutions. The MALL and CUHK datasets in Table 1 were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

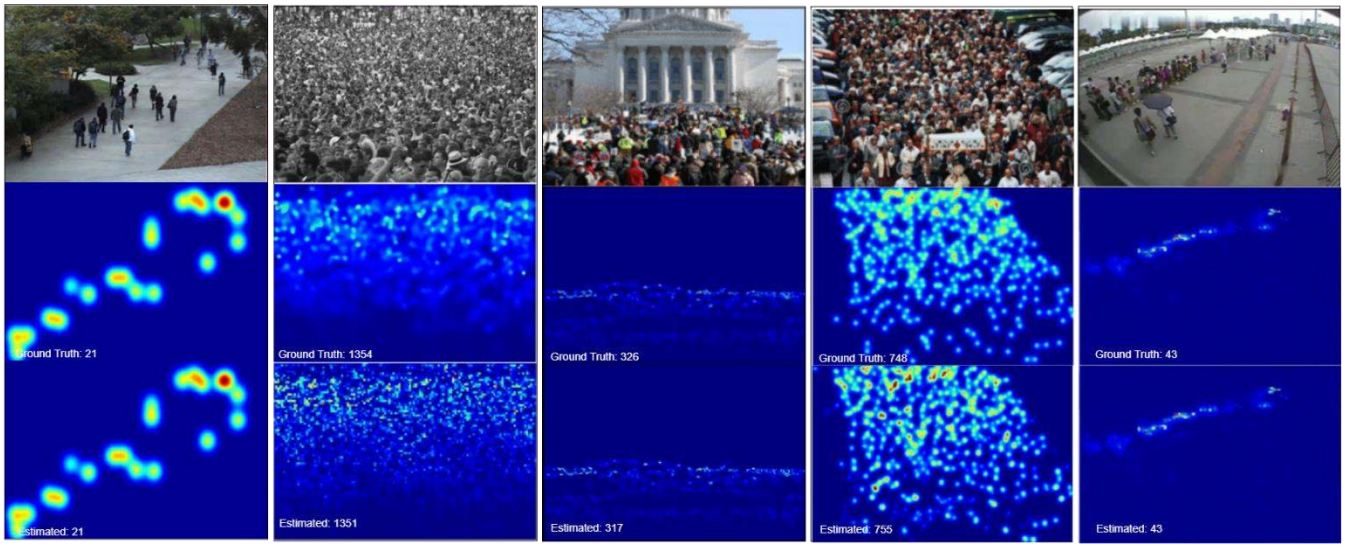

FIGURE 4. Qualitative results of the ground truth and estimated crowd count are shown respectively column from left to right on the datasets UCF-QNRF, UCF_CC_50, UCSD, ShangaiTech Part A, and WorldExpo'10.

not used in the study because the MALL consists only of images indoors and the CUHK relates to crowd behavior.

B. EVALUATION METRICS

By using the proposed model, the number of people is estimated more effectively from the density map. For this purpose, alpha (σ) and delta (δ) parameters of Gaussian cores are used as in equation (16) to calculate the ground truth of the density map.

$$Y = \delta(c - c_i) * G_\sigma, (Gaussian\ kernel\ \sigma = 4) \quad (16)$$

For pre-trained base model training, batch size 16, momentum 0.9, weight decay L2 penalty 5e-4, optimization algorithm Adam, and learning rate 1e-4 are selected.

Model performance is evaluated by two metrics, Mean Absolute Error (MAE) and Mean Square Error (MSE), and compared with similar studies in the literature.

The MAE reflects the accuracy of the results predicted by the model, and the MSE is used to measure the robustness of the model. It is defined as equations (17) and (18) below, where N is the number of images in the test set, i is the index of the image, y_i is the estimated number calculated for the image, and y_i^{GT} is the number of ground truth in the image.

$$MAE = \frac{1}{N}\sum_{i=0}^{N}\left|y_i - y_i^{GT}\right| \quad (17)$$

$$MSE = \frac{1}{N}\sum_{i=0}^{N}\left|y_i - y_i^{GT}\right|^2 \quad (18)$$

A Structural Similarity Index (SSIM) comparison is also made for the ShangaiTech Part A dataset. Thus, the quality of the density map is also measured and defined as in equation (19).

$$SSIM = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{(2\mu P_{yi}.\mu G_{yi}GT+y_1)}{\mu_{P_{yi}}^2+\mu_{G_{yi}}^2+y_1}\cdot\frac{(2\sigma P_{yi}.\sigma G_{yi}GT+y_2)}{\sigma_{P_{yi}}^2+\sigma_{G_{yi}}^2+y_2}\right) \quad (19)$$

Where, μP_{yi} and μG_{yi}mean, σP_{yi}, σG_{yi}, and $\sigma P_{yi}.\sigma G_{yi}GT$ standard deviation values are used for SSIM loss calculation.

C. EXPERIMENTAL RESULTS

Evaluation metrics are compared with recent studies in order to comprehensively evaluate the performance of the proposed approach. For this, experiments are carried out on five datasets.

- In the UCF-QNRF [72] dataset, the proposed method yielded a 2.18% increase in accuracy over the next best result for MSE [80].
- In the UCF_CC_50 [74] dataset, a 2.33% more accurate MAE result is obtained than in [84] and a 1.68% more exact MSE result is obtained compared with [80].
- In the UCSD [75] dataset, improved accuracy of 8.46% for MAE result compared with [36] and a 7.03% more competitive MSE result is obtained compared with [49].
- In the ShanghaiTech Part A [76] dataset, more competitive results are obtained compared to [29], with improved accuracy of 4.69% MAE and 3.94% MSE.
- In the WorldExpo'10 [77] dataset, MAE is 6.94% better than the closest result [80].

Experimental results for these five datasets are compared in Tables 2 to 7, and Figure 4 shows the comparison of estimates obtained from density maps of the UCF-QNRF, UCF_CC_50, UCSD, ShanghaiTech Part A, and WorldExpo'10 datasets. Table 6 compares the quality of the density map estimated for the ShangaiTech Part A dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

TABLE 2. Comparison results on the UCF-QNRF.

Method	Estimation Error of Crowd Count	
	MAE	MSE
Sam *et al.* [16]	252	514
Idrees *et al.* [74]	277	426
Badrinarayanan *et al.* [82]	228	445
He *et al.* [23]	190	277
Huang *et al.* [83]	163	226
Idrees *et al.* [81]	132	191
Liu *et al.* [80]	107	183
Proposed Method	109	179

TABLE 3. Comparison results on the UCF_CC_50.

Method	Estimation Error of Crowd Count	
	MAE	MSE
Kang *et al.* [30]	406.2	404
Zhang *et al.* [15]	377.6	509.1
Sam *et al.* [16]	318.1	439.2
Zhu *et al.* [19]	302.3	411.6
Shi *et al.* [18]	288.4	404.7
Li *et al.* [32]	266.1	397.5
Basalamah *et al.* [49]	235.74	345.6
Liu *et al.* [80]	212.2	243.7
Xu *et al.* [84]	188.4	315.3
Proposed Method	184	239.6

TABLE 4. Comparison results on the UCSD.

Method	Estimation Error of Crowd Count	
	MAE	MSE
Zhang *et al.* [15]	1.60	3.31
Sam *et al.* [16]	1.62	2.10
Kang *et al.* [30]	1.17	2.15
Li *et al.* [32]	1.16	1.47
Liu *et al.* [80]	1.03	1.32
Basalamah *et al.* [49]	1.01	1.28
Jiang *et al.* [36]	0.97	1.30
Proposed Method	0.89	1.19

TABLE 5. Comparison results on the ShangaiTech Part A.

Method	Estimation Error of Crowd Count	
	MAE	MSE
Zhang *et al.* [14]	181.8	277.7
Marsden *et al.* [79]	126.5	173.5
Zhang *et al.* [15]	110.2	173.2
Sam *et al.* [16]	90.4	135.0
Chen *et al.* [73]	73.6	106.4
Gao *et al.* [41]	73.5	124.0
Li *et al.* [32]	68.2	115.0
Shi *et al.* [33]	66.3	106.4
Jiang *et al.* [36]	62.4	102.6
Liu *et al.* [80]	62.0	100.0
Sindagi *et al.* [29]	46.9	71.0
Proposed Method	44.7	68.2

TABLE 6. Comparison results for quality of density map on the ShangaiTech Part A.

Method	Quality of Density Map
	SSIM
Zhang *et al.* [15]	0.52
Chen *et al.* [73]	0.72
Li *et al* [32]	0.76
Proposed Method	0.43

TABLE 7. Comparison results on the WorldExpo'10.

Method	Estimation Error of Crowd Count
	MAE
Zhang *et al.* [14]	12.9
Zhang *et al.* [15]	11.6
Sam *et al.* [16]	9.4
Shi *et al.* [18]	9.1
Sindagi *et al.* [17]	8.86
Li *et al.* [32]	8.6
Shen *et al.* [44]	7.5
Basalamah *et al.* [49]	7.42
Liu *et al.* [80]	7.2
Proposed Method	6.7

1) ABLATION EXPERIMENTS ON DENSITY MAP PARAMETERS

In this section, Gaussian core value $\sigma = 4$ is chosen when creating density maps in crowd counting tasks. In this experiment, when σ value 16 is selected, the results are compared in all datasets. It is shown that the best result is achieved with the proposed method. Results of the value of the Gauss kernel ablation experiment for these three datasets are compared in Table 8.

TABLE 8. Results of the value of Gauss kernel ablation for density map generation.

Datasets	Value of Gauss Kernel	MAE	MSE
UCF-QNRF [72]	σ =4	109	179
	σ =16	112.2	187.1
UCF_CC_50 [74]	σ =4	184	239.6
	σ =16	187.2	242
UCSD [75]	σ =4	0.89	1.19
	σ =16	0.93	1.32
ShangaiTech Part A [76]	σ =4	44.7	68.2
	σ =16	47.2	71.0
WorldExpo'10 [77]	σ =4	6.7	-
	σ =16	7.2	-

2) ABLATION EXPERIMENTS ON THE STRUCTURE OF THE PROPOSED METHOD

In this section, ablation studies are performed to validate the multi-level spatial information of the CNNs and CapsNet attention modules. For the datasets used, the experiments are repeated with the convolution attention module only and the CapsNet attention module only. Results are compared in Table 9. In the crowd analysis of this model architecture, it is shown that it predicts the number of people successfully with more robust and better density map quality than in many recent studies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3081529, IEEE Access

TABLE 9. Results of the ablation model structures on the datasets UCF-QNRF, UCF_CC_50, UCSD, ShangaiTech Part A and WorldExpo'10.

Datasets	Method	Modules		
		Base + Convolution Attention + Compress.	Base + Capsule Attention + Compress.	Proposed Method
UCF-QNRF [72]	MAE	143	110	109
	MSE	198	182	179
UCF_CC_50 [74]	MAE	225.6	206	184
	MSE	287	244.3	239.6
UCSD [75]	MAE	1.23	1.03	0.89
	MSE	1.55	1.20	1.19
ShangaiTech Part A [76]	MAE	65.6	59.6	44.7
	MSE	102.8	98.1	68.2
WorldExpo'10 [77]	MAE	8.8	7.6	6.7

V. CONCLUSION AND DISCUSSION

In this paper, we use the spatial information learning feature of Capsule networks to estimate the number of people, especially in crowded images. The VGG-16 network was used to extract basic-level features. We put forward the first crowd analysis study using a two-column cascade and CNN and CapsNet as an attention module. The positive effect of the Capsule attention module in this study is emphasized in the ablation study for structure. In order to determine the number of people in dense crowd images, features such as posture and angular value have achieved more robust results with the capsule attention mechanism. This propounded strategy has been tested on the UCF-QNRF, UCF_CC_50, UCSD, ShanghaiTech Part A, and WorldExpo'10 datasets. The performance of the proposed approach is shown to be effective for this problem when compared to state-of-the-art approaches.

The method we propose is still not good in terms of computational complexity. In the near future, we plan to improve computational complexity by using model lightweight technology and present results in comparison with those of state-of-the-art products. We anticipate that CapsNet's position and orientation information can also be used for crowd behavior analysis without using motion information such as optical flow in future studies.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research.

REFERENCES

[1] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object detection, tracking and recognition for multiple smart cameras," in *Proc. IEEE*, vol. 96, no. 10, 2008, pp. 1606–1624.

[2] S. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, Mar. 2009.

[3] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008.

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[5] N. Anjum and A. Cavallaro, "Trajectory association and fusion across partially overlapping cameras," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2009, pp. 201–206.

[6] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Jun. 2016, pp. 4256–4265.

[7] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[8] L. D. Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Counting people by RGB or depth overhead cameras," *Pattern Recognit. Lett.,* vol. 81, pp. 41–50, Oct. 2016.

[9] C. Ertler, H. Possegger, M. Opitz, and H. Bischof, "Pedestrian detection in RGB-D images from an elevated viewpoint," *in Proc. 22nd Comput. Vis. Winter Workshop*, 2017, pp. 1–9.

[10] I. Ahmed and A. Adnan, "A robust algorithm for detecting people in overhead views," *Cluster Comput.*, vol. 21, no. 1, pp. 633–654, Mar. 2018.

[11] P. Baque, F. Fleuret, and P. Fua, "Deep occlusion reasoning for multicamera multi-target detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 271–279.

[12] J. Wetzel, A. Laubenheimer, M. Heizmann, "Joint Probabilistic People Detection in Overlapping Depth Images," *IEEE Access*, vol. 8, pp. 28349 - 28359, 2020.

[13] T. Chavdarova and F. Fleuret, "Deep multi-camera people detection," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 848–853.

[14] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[15] 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. (CVPR)*, Jun. 2016, pp. 589–597.

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

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

[18] 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.

[19] L. Zhu, C. Li, Z. Yang, K. Yuan, and S. Wang, "Crowd density estimation based on classification activation map and patch density level," in *Neural Computing & Applications. Berlin, Germany: Springer*, 2019, pp. 5105–5116.

[20] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *ACM International Conference on Multimedia*, pp. 1299–1302, 2015.

[21] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Engineering Applications of Artificial Intelligence*, vol. 43, pp. 81–88, 2015.

[22] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 833–841.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* 2016.

[24] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," *in IEEE Int. Conf. on Image Process.* pp. 1215–1219, 2016.

[25] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," *in ACM International Conference on Multimedia*, 2016, pp. 640–644.

[26] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *European Conference on Computer Vision*, 2016, pp. 483–498.

[27] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *European Conference on Computer Vision*, 2016, pp. 785–800.

[28] Z. Zhao, H. Li, R. Zhao, and X. Wang, "Crossing-line crowd counting with two-phase deep neural networks," in *European Conference on Computer Vision*, 2016, pp. 712–726.

[29] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance*, Italy, 2017.

[30] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks - counting, detection, and tracking," in *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 99, May 2019,

[31] N. Ilyas, A. Ahmad, K. Kim, "CASA-Crowd: A Context-Aware Scale Aggregation CNN-Based Crowd Counting Technique," in *IEEE Access*, vol. 7, pp. 182050 - 182059, 2019.

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

[33] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7279–7288.

[34] 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.

[35] 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.

[36] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li, Y. Pang, X. Li, B. Zhou, and M. Xu, "Learning multi-level density maps for crowd counting," in *IEEE Trans. Neural Netw. Learn. Syst*., vol. 31, no. 8, pp. 2705–2715, Aug. 2020.

[37] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Springer, Amsterdam, Netherlands, 2016, pp. 615–629.

[38] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, p. 7, 2017.

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

[40] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "PaDNet: Pan-density crowd counting," in *IEEE Trans. on Image Proces.*, vol. 29, pp 2714 – 2727, 2019.

[41] J. Gao, Q. Wang, and X. Li, "PCC Net: Perspective crowd counting via spatial convolutional network," in *IEEE Trans. Circuits Syst. Video Technol*., vol. 30, no. 10, pp. 3486–3498, Oct. 2020.

[42] 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.

[43] D. Guo, K. Li, Z.-J. Zha, and M. Wang, "DADNet: dilated-attention deformable ConvNet for crowd counting," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1823–1832.

[44] 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.

[45] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5197–5206.

[46] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Person head detection in multiple scales using deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

[47] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Crowd counting in low-resolution crowded scenes using region-based deep convolutional neural networks," in *IEEE Access*, vol. 7, pp. 35317–35329, 2019.

[48] M. Shami, S. Maqbool, H. Sajid, Y. Ayaz, and S.-C. S. Cheung, "People counting in dense crowd images using sparse head detections," in *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2627- 2636, 2019.

[49] S. Basalamah, S. D. Khan, H. Ullah, "Scale Driven Convolutional Neural Network Model for People Counting and Localization in Crowd Scenes," in *IEEE Access*, vol. 7, pp. 71576 - 71584, 2019.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Jun. 2018, pp. 7132–7141.

[51] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.

[52] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, arXiv:1807.06514. [Online]. Available: http://arxiv.org/abs/1807.06514.

[53] J. U. Kim and Y. Man Ro, "Attentive layer separation for object classification and object localization in object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3995–3999.

[54] M. Shaikh, V. A. Kollerathu, and G. Krishnamurthi, "Recurrent attention mechanism networks for enhanced classification of biomedical images," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1260–1264.

[55] Z. Lan, Q. Huang, F. Chen, and Y. Meng, "Aerial image semantic segmentation using spatial and channel attention," in *Proc. IEEE 4th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2019, pp. 316–320.

[56] H. Ling, J. Wu, L. Wu, J. Huang, J. Chen, and P. Li, "Self-residual attention network for deep face recognition," in *IEEE Access*, vol. 7, pp. 55159–55168, 2019.

[57] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Jun. 2018, pp. 7794–7803.

[58] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[59] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[60] 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.

[61] Z. Huo, B. Lu, A. Mi, F. Luo, Y. Qiao, "Learning multi-level features to improve crowd counting," in *IEEE Access*, vol. 8, pp. 211391 - 2114006, 2020.

[62] H. H. Nguyen, J. Yamagishi, and I, Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP),* 2301–2307 (2019).

[63] A. M. Algamdi, V. Sanchez, and C. T. Li, "Learning temporal information from spatial information using CapsNets for human action recognition" in *IEEE Int. Conf. Acoust. Speech Signal Process*, pp 3867–3871 (2019).

[64] S. Liu, W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, doi:10.1109/ACPR.2015.7486599, 2015.

[65] D. H. Hubel, and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," in *The Journal of Physiolog*y, vol. 195, no. 1, pp. 215-43, 1968.

[66] K. N. Fukushima, "A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," in *Biol. Cybern*., vol. 36, no. 4, pp. 193–202, 1980.

[67] Y. LeCun, C. Cortes, and J. B. Burges, "The MNIST database of handwritten digits," 1998, [Online]. Available: http://yann.lecun.com/exdb/mnist/

[68] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. 31st Conf. on Neural Information Proces. Systems (NIPS)*, 2017, pp. 3859–3869.

[69] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming autoencoders," in *Int. Conf. on Artificial Neural Networks (ICANN),* Springer, Berlin, Heidelberg, 2011, pp. 44-51.

[70] W. Huang, and F. Zhou, "DA-CapsNet: dual attention mechanism capsule network," Nature Scientific Reports vol. 10, art. num.: 11383, 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-68453-w

[71] A. Khan, J. A. Shah, K. Kadir, W. Albattah, and F. Khan, "Crowd monitoring and localization using deep convolutional neural network: A review," in *Appl. Sci*., vol. 10, no. 14, pp. 4781, 2020.

[72] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 8–14 September, 2018, pp. 532–546.

[73] K. Chen, S. Gong, T. Xiang, C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, OR, USA, 23–28 June, 2013, pp. 2467–2474.

[74] H. Idrees, I. Saleemi, C. Seibert, M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, OR, USA, 23-28 June 2013, pp. 2547–2554.

[75] A. B. Chan, Z. S. J. Liang, N. Vasconcelos, "Privacy-preserving crowd monitoring: counting people without people models or tracking," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 23-28 June, 2008, pp. 1–7.

[76] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June, 2016.

[77] C. Zhang, H. Li, X. Wang, X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Boston, MA, USA, 7–12 June, 2015.

[78] J. Shao, K. Kang, C. C. Loy, X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, 7–12 Jun, 2015, pp. 4657–4666.

[79] M. Marsden, K. McGuiness, S. Little, and N. O'Connor, "Fully convolutional crowd counting on highly congested scenes," in *Proc. of the 12th Inter. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP),* 2017, pp. 27-33.

[80] W. Liu, M. Salzmann, P. Fua, "Context-aware crowd counting," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, CA, USA, 2019.

[81] 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 European Conference on Computer Vision*, Munich, Germany, 8-14 Sep., 2018, pp.544-559.

[82] V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, Jan. 2017.

[83] G. Huang, Z. Liu, L. V. Maaten, K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., USA, 21–26 July 2017, pp. 2261 – 2269.

[84] C. Xu, K. Qiu, J. Fu, S. Bai, Y. Xu, X. Bai, "Learn to scale: generating multipolar normalized density maps for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis,* Seoul, Korea, 2019, pp. 8382–8390.

MERVE AYYÜCE KIZRAK was born in Istanbul, Turkey. She was awarded B.S. and M.S. degrees in Electronics and Communication Engineering from Haliç University in 2009 and 2011, respectively. She is a Ph.D. candidate in Electronics and Communication Engineering at Yıldız Technical University in Istanbul, Turkey.

From 2009 to 2019, she was a Research Assistant with the Electrical and Electronics Engineering Department, Haliç University. Since 2019 she has been a Research Assistant with the Artificial Intelligence Engineering Department, Bahçeşehir University. She has been employed as an AI expert at The Digital Transformation Office, The Presidency of the Republic of Turkey since 2019. Her research interests include image processing, computer vision application to crowd analysis, and machine learning.

She was the chair of the IEEE student branch at Haliç University between 2007 and 2011. Since 2018, she has been engaged in volunteer work in the field of AI as a Turkey deeplearning.ai ambassador.

BÜLENT BOLAT was born in Kahramanmaraş, Turkey. He received his BS, MSc and PhD degrees from Yıldız Technical University, Faculty of Electrical and Electronics Engineering, Electronics and Communications Engineering Department. During his MSc and PhD period, he worked as a Research Assistant in the same department. He is still working as an Associate Professor at Yıldız Technical University. He also worked as guest lecturer for several universities in Turkey.