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. 2017. Doi Number

Learning Multi-level Features to Improve Crowd Counting

Zhanqiang Huo[1], Bin Lu[1], Aizhong Mi[1], Fen Luo[1], and Yingxu Qiao[1]

[1]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454003, China

Corresponding author: Fen Luo (e-mail: luofenjsj@hpu.edu.cn).

This work is supported in part by the Henan University Scientific and Technological Innovation Team Support Program (19IRTSTHN012) and in part by the Foundation of Henan Scientific and Technological Project (192102210281).

ABSTRACT Crowd counting is a task that aims to estimate the number of people in an image. Recent crowd counting methods make significant progress by employing convolutional neural networks to regress crowd density maps. One of the most challenging problems in this task is the drastic scale variation of the region of interest in images. In this paper, a Feature Fusion Attention Network (FFANet) is proposed for crowd counting. Firstly, the VGG16 network is adapted as the backbone of the FFANet to extract the features of crowd images. Then, the extracted features are fused by the subsequent two stages. Specifically, the information enhancement operations on the multi-levels features are conducted by Feature Fusion Attention Module (FFAM), which are further refined by the Residual Block (RB). Finally, the features are processed by the Compression Module (CM) to generate a density map. To demonstrate the effectiveness, the proposed algorithm is verified on three benchmark datasets. Evaluation of the algorithm performances in comparison with other state-of-the-art methods indicates the proposed FFANet outperforms the existing methods.

INDEX TERMS Crowd Counting, Scale Variation, Feature Fusion Attention.

I. INTRODUCTION

Crowd counting is one of the promising applications in computer vision. It is a task that aims to estimate people's numbers in an image. The predicted results can be used in a wide range of fields, such as intelligent transportation [1], public security [2], agriculture monitoring [3], video surveillance [4] and so on. However, crowd counting is also a highly challenging task because of occlusion, low image resolution, perspective distortion, scale variation of objects, etc [5]. To obtain accurate results, researchers have paid lots of attention to study the above issues.

Early methods for crowd counting are based on manual features extraction of the human body and various regression functions [6]. These methods usually don't perform well in dense crowd scenes where pedestrians are severely occluded or overlapped. With the development of deep learning technology in computer vision, the algorithms based on CNN have made great progress by conquering the scales variation. These methods typically design different sized filters architectures to extract multi-scale features [7]. However, human scales change continuously in the entire image and current models can only concern some discrete scales. It brings a major problem that these methods ignore a larger number of crowds in an image.

Some researchers [8]-[10] discover that CNN's shallower layers focus on low-level texture and spatial information which can help the model determine the location of the target. The deeper layers focus on high-level contextual and semantic information which can help the model to identify the type of targets. Inspired by the above research, we reasoned that the fusion of these multi-level features can effectively solve the crowd scales variation. There are two ways to realize feature fusion, one is to merge the features on its channel axis and the other is to conduct element-wise summation. The defect of these two methods is that they can't utilize the information contained in the feature effectively, which results in the waste of calculation. In this paper, we introduce the FFAM to realize the information enhancement of the multi-level features. First, the FFAM utilizes the contextual and semantic information contained in the high-level features to enhance channel-wise information in the low-level features. Second, the spatial information of the processed low-level features would be extracted by the FFAM to enhance the high-level features. Third, the features

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.2020.3039998, IEEE Access

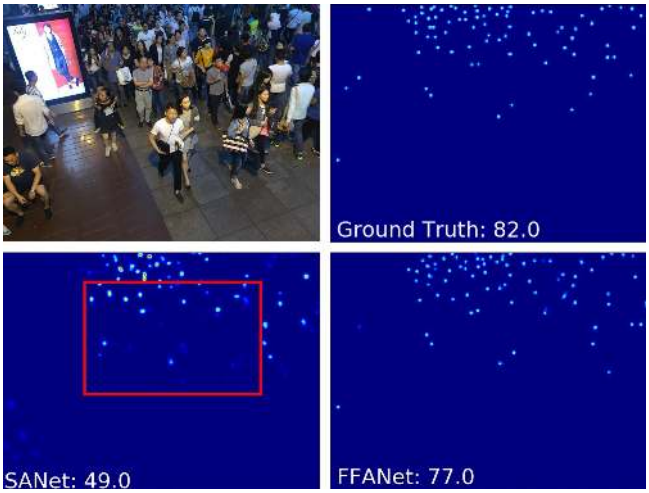

FIGURE 1. Density estimation results. Top left: input image. Top right: ground truth. Bottom left: SANet [9]. Bottom right: our FFANet.

processed in the first two steps are concatenated in the channel axis. The experiment shows that FFAM significantly improves the accuracy of crowd counting tasks.

Fig. 1 shows a crowd image and the estimated density maps by the proposed FFANet and the SANet [11]. Compared with SANet, our result deviate less from the ground truth. By observing the spatial distribution of the density maps, SANet could not solve the drastic scale variation of the region of interest in crowd images (as shown in the red box). On the contrary, the proposed FFANet solves the scale variation problem well and the spatial distribution of the estimated density map is very similar to the ground truth. In conclusion, compared with SANet, the FFANet proposed in this paper has a significant improvement in solving the problem of the drastic scale variation in crowd image and improving the accuracy of crowd counting.

In summary, the contributions of this paper are as follows:

1) A new end-to-end multi-level feature fusion network is proposed to enhance the network robustness to scale variations of crowd images.
2) The proposed FFAM is utilized to enhance the spatial and semantic information between multi-level features in the FFANet.

The remainder of the paper is organized as follows. After the related work discussion in Section II, we cover the details of our proposed method in Section III. Section IV introduces experimental designs and discusses the results. We conclude with a short discussion in Section V.

II. RELATED WORK

A. CROWD COUNTING

In recent years, crowd counting methods have made great progress by employing convolutional neural networks to regress crowd density maps. Researchers have designed a variety of efficient convolution neural networks to solve scale variation [5]. The remainder part of this section describes the multi-column models and the single-column models according to the network structure.

Multi-column models: Zhang *et al.* [7] proposed a Multi-column Convolutional Neural Network (MCNN) to overcome the scale variation in images. Each column is composed of different filters to get features with various scales. Inspired by MCNN, Onoro *et al.* [12] designed a scale-aware counting model that can predict crowd distribution and the number of the crowd without perspective information, by extracting features from the image with different resolutions to overcome the perspective distortion. Switch-CNN which trains a classifier to choose the optimal branch from the multi-column network for crowd image patches is proposed by Sam *et al.* [13]. SANet designed by Cao *et al.* [11] uses scale aggregation modules to solve scale variation and extracts these features to generate high-resolution density maps. Guo *et al.* [14] explored a scale-aware attention fusion with various dilation rates to capture different visual granularities of crowd regions of interest and utilizes deformable convolutions to generate a high-quality density map. Recently, Gao *et al.* [15] proposed a Perspective Crowd Counting via Spatial Convolutional Network (PCC Net) to solve high appearance similarity, perspective changes and severe congestion.

Single-column models: CSRNet [16] used dilated convolution layers to expand the receptive field and replace pooling operations. By taking advantage of these designs, CSRNet can easily generate high-quality density maps. Shi *et al.* [17] proposed a Perspective Aware Convolutional Neural Network (PACNN), which can add perspective information based on crowd estimation and effectively solve the scale variation. ADCrowdNet [18] consists of two CNN networks. An attention aware network firstly detects the crowd regions in the image and calculates the congestion degree of these areas. Based on the detected crowd area, a multi-scale deformable network is used to generate high-quality density maps. More recently, Jiang *et al.* [19] proposed an effective Multi-Level Convolutional Neural Network (MLCNN) architecture that first adaptively learns multi-level density maps and then fuses them to predict the final output.

B. ATTENTION MODELS

Attention models are first applied in the field of machine translation and then developed into many deep learning fields such as object detection [20], image classification [21], image segmentation [22] and face recognition [23]. Hu *et al.* [24] proposed a lightweight attention mechanism named SENet which could automatically obtain the importance of each feature channel-wise information to enhance the useful features. Jon. *et al.* combined the channel attention mechanism and the spatial attention mechanism to propose CBAM [25] and BAM [26]. Li *et al.* [27] proposed a dynamic selection mechanism that allows

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.2020.3039998, IEEE Access

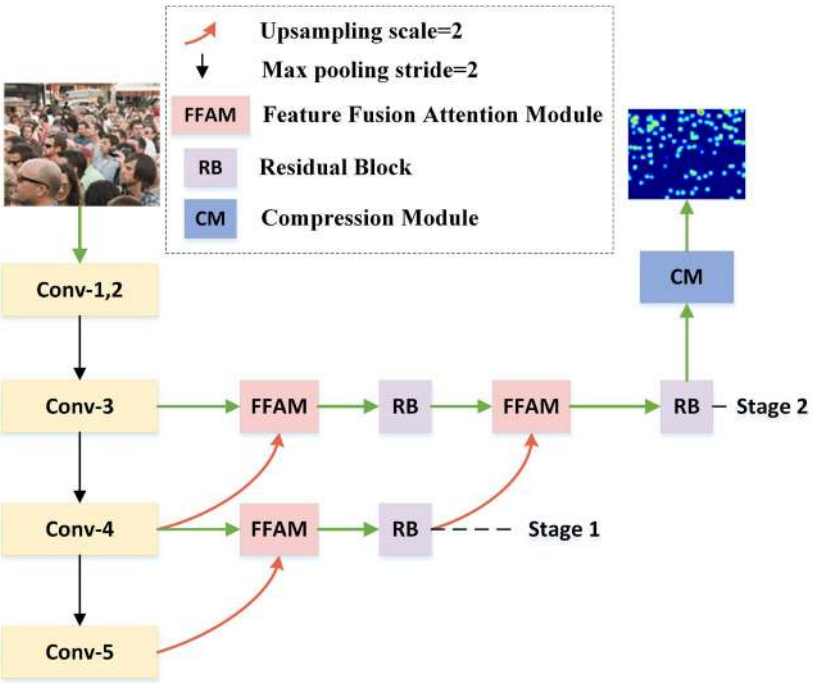

FIGURE 2. **Overview of the proposed FFANet. The image is fed to the first five convolutional blocks of VGG16 for extracting features. The output from the conv_5 is upsampled and fused with the output from conv_4 to achieve one-stage feature fusion. Then the fused feature maps and the output of the conv_4 are further upsampled and fused with the feature maps from conv_3 to achieve two-stage feature fusion. Finally, the final fused features are processed by the CM to generate a density map.**

each neuron to select a different receptive field for the size of the target. Cao et al. [28] combined the advantages of Non-local [29] and SENet and proposed a new global context module, achieving significant results on computer vision tasks.

Multi-column models typically design different sized filters architectures to extract multi-scale features. However, human scales change continuously in the entire image and current models can only concern some discrete scales. It brings a major problem that these methods ignore a larger number of crowds in an image. Single-column models need to rely on other information or auxiliary network to solve the scale variation, for example, PACNN needs to generate perspective maps and train the corresponding network branches. This has led to additional work and increased difficulty in model training. Therefore, combined with the above works, we design a simple and effective network based on multi-level feature fusion to solve the problem of scale variation in crowd counting.

III. METHOD

In this section, we firstly introduce the structure of the proposed FFANet. Then, we introduce the associated modules and the loss function. Fig. 2 shows the details of the network structure. Table I describes the configuration of convolutional layers in the proposed FFANet. [(3,3)-64-BN-ReLU] means that the convolutional block contains 3×3 kernel size convolutional operation, 64 output channels, a BN [30] layer and a ReLU activation layer. (FFAM+RB) represents the cascade structure composed of the FFAM and the RB.

TABLE I. Configuration of convolution layers in the proposed FFANet.

FFANet
Backbone
Conv-1: [(3,3) - 64 - BN - ReLU] × 2
Max pooling
Conv-2: [(3,3) - 128 - BN - ReLU] × 2
Max pooling
Conv-3: [(3,3) - 256 - BN - ReLU] × 3
Max pooling
Conv-4: [(3,3) - 512 - BN - ReLU] × 3
Max pooling
Conv-5: [(3,3) - 512 - BN - ReLU] × 3
Max pooling
Stage 1: (FFAM+ RB) × 1
Upsample 5 level Features
FFAM × 1
RB: [(1,1) - 256 - ReLU] × 1
[(3,3) - 256 - BN – ReLU] × 1
[(3,3) - 256 - ReLU] × 1
Stage 2: (FFAM+ RB) × 2
Upsample 4 level Features
FFAM × 1
RB: [(1,1) - 128 - ReLU] × 1
[(3,3) - 128 - BN - ReLU] × 1
[(3,3) - 128 - ReLU] × 1
Compression Module
Conv-1 : [(3,3) - 32 - ReLU] × 1
Conv-2 : [(1,1) - 1 - ReLU] × 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.2020.3039998, IEEE Access

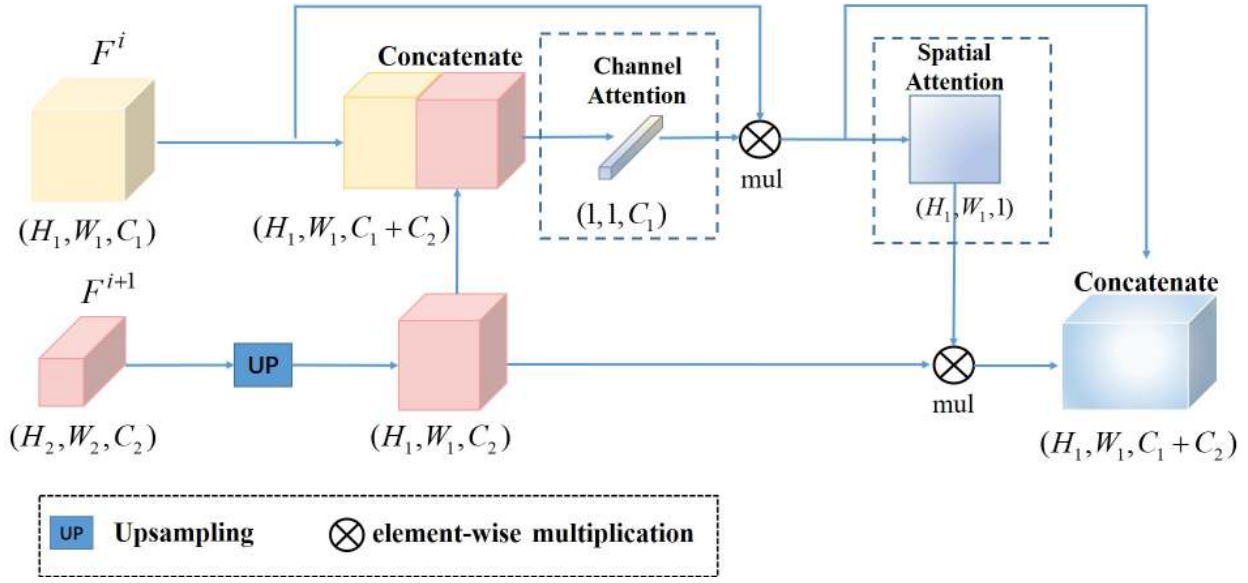

FIGURE 3. The architecture of the proposed FFAM. This F^i indicates the feature maps output from the *i*-th convolutional block. The concatenate means to concatenate the feature maps in the channel axis. (H_1, W_1, C_1) represents the shape of the feature.

A. OVERVIEW

The FFANet consists of a backbone network, feature fusion attention modules, residual blocks and a compression module.

Backbone the backbone network is a pre-trained VGG16 [31] network with the fully connected layers removed. A BN layer is added at the back of all convolutional layers in the VGG16 network. The backbone network contains 13 convolutional layers which are divided into 5 convolutional blocks. We take the output of convolution blocks from the third to the fifth as the objects of feature fusion.

FFAM the proposed FFAM fuses multi-level features of the backbone and utilizes this information diversity to enhance channel-wise information in the low-level feature and spatial-wise information in the high-level feature.

RB & CM the RB is a residual block composed of a 1×1 kernel size convolutional layer and two 3×3 convolutional layers to refine the features of FFAM output. The compression module (CM) compresses the feature map into a single-channel crowd density map.

B. FEATURE FUSION ATTENTION MODULE

Fig. 3 displays the architecture of the FFAM. The input is the features of two adjacent levels in the backbone network. The FFAM upsamples the high-level feature F^{i+1} and concatenates it with the low-level feature F^i in the channel axis. The channel attention module uses the concatenated features to output vectors w^c to enhance channel-wise information in F^i. Fig. 4 (a) describes the structure of the channel attention module. It is formulated as

$$w^c = \phi([F^i, U(F^{i+1})]) \tag{1}$$

where $U(\cdot)$ denotes the upsampling layer, $[\cdot]$ denotes the concatenation layer, $\phi(\cdot)$ denotes the channel attention module. Channel-wise enhanced F' is obtained by element-wise multiplying vector w^c and F^i. It is formulated as

$$F' = w^c \otimes F^i \tag{2}$$

The spatial attention module calculates F' to obtain the spatial weight to enhance the high-level feature. Fig. 4 (b) represents the structure of the spatial attention module. We define this operation as

$$w^s = \varphi(F') \tag{3}$$

where $\varphi(\cdot)$ denotes the spatial attention module, w^s denotes the spatial weight. Spatial-wise enhanced F'' is obtained by element-wise multiplying w^s and $U(F^{i+1})$. It is formulated as

$$F'' = w^s \otimes U(F^{i+1}) \tag{4}$$

Finally, the two enhanced features would be concatenated in the channel axis. It is formulated as

$$\hat{F} = [F', F''] \tag{5}$$

C. LOSS FUNCTION

We define a joint loss function which consists of Mean Squared Error (MSE) loss and Structural Similarity Index (SSIM) loss. MSE loss is used to minimize the Euclidean distance between the ground truth and the estimated density map. Ref. [11] reveals the fact that MSE loss employed by many previous methods is dependent on the pixel independence hypothesis and doesn't consider the local correlation of the density map. Therefore, we utilize the SSIM loss as part of the loss function to improve the result. The joint loss function is defined as follows

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.2020.3039998, IEEE Access

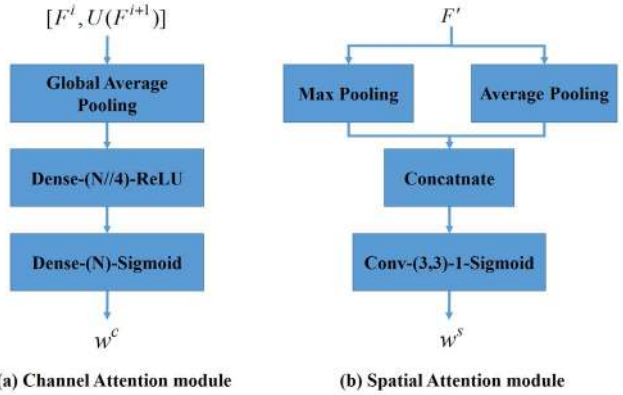

FIGURE 4. Structure of the Channel Attention module and Spatial Attention module. Dense-(N)-Sigmoid means that the dense layer contains N neurons and the Sigmoid activation function. The N of the dense layer is set to {128, 256} in stages 1, 2 in sequence. Conv-(3, 3)-1-Sigmoid means that the convolutional layer contains a 3×3 kernel size convolutional operation, 1 output channels and the Sigmoid activation function.

$$
\begin{aligned}
L &= L^{MSE} + \lambda L^{SSIM} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\| D_i^G - D_i^P \right\|_2^2 \\
&+ \lambda \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{1}{M}\sum_{j} SSIM_i(j)) \\
SSIM_i &= \frac{(2\mu_{P_i}\mu_{G_i} + C_1)}{\mu_{P_i}^2 + \mu_{G_i}^2 + C_1} \cdot \frac{(2\sigma_{P_iG_i} + C_2)}{\sigma_{P_i}^2 + \sigma_{G_i}^2 + C_2}
\end{aligned}
\tag{6}
$$

where N is the number of training batch size, D_i^G is the ground truth of density map, D_i^P is the predicted density map, M is the number of the pixels in the density map, λ is the parameters which are used to balance L^{MSE} and L^{SSIM}. Means (μ_{P_i}, μ_{G_i}) and standard deviation $(\sigma_{P_i}, \sigma_{G_i}, \sigma_{P_iG_i})$ in SSIM loss are calculated with a Gaussian kernel that configures a standard deviation of 1.5 within an 11×11 region at each position j.

IV. EXPERIMENT

In this section, we validate the effectiveness of our method on three public crowd counting datasets: ShanghaiTech [7], UCF_CC_50 [32], UCSD [33]. Then we conduct the ablation studies about the hyperparameter λ, the density map generation and the structure of the network.

A. DATASETS AND TRAINING DETAILS

- ShanghaiTech. The ShanghaiTech dataset includes 1198 images with 330,165 annotated heads. It consists of two parts: Part A and Part B. Specifically, Part A consists of 482 images, which are randomly selected from the Internet and Part B is selected from the surveillance on the streets of Shanghai. These two parts are further divided into training and evaluation sets. 300 and 182 images from Part A are selected for training and testing respectively, while 400 and 316 images from Part B are also chosen.
- UCF_CC_50. The UCF_CC_50 dataset contains 50 images with different resolutions and each image has an average of 1280 people. In the whole dataset, the number of individuals in each image ranges from 94 to 4543, which indicates that there is a large count variance between images. Considering the small size of the dataset, we use a cross-validation protocol for training and testing our methods following the approach from ref. [7].
- UCSD. The dataset consists of 2000 frames from a surveillance video camera on the UCSD campus. The resolution of each frame is 158 × 238. The average number of people in each frame is 25. The dataset provides the region of interest to ignore the background. Following ref. [11], frames #601 to #1400 are used for training and the rest for testing. To satisfy the constraints of the backbone on the shape of the input tensor, we resize the resolution of the image to 400× 608. This operation can not only meet the input restrictions but also ensure that the image content is not distorted.

If the image in the dataset has the various resolution, the original image will be cropped into 400 × 400 patches. We convert the label to a density map by a fixed-size Gaussian kernel $G_\sigma(\sigma = 4)$. We use the delta function $\delta(x - x_i)$ to represent the head position. The ground truth of density map Y is generated by convolving the Gaussian kernels with each delta function.

$$Y = \delta(x - x_i) * G_\sigma \tag{7}$$

The parameters of the network are randomly initialized by the Gaussian distribution with mean zero and standard deviation of 0.01 except for the backbone network. We use Adam with a learning rate of 1e-4 and a batch size of 16 to train the network.

B. EVALUATION METRICS

The performance of the model is evaluated by two metrics, the Mean Absolute Error (MAE) and Mean Square Error (MSE). MAE reflects the accuracy of the results predicted by the model and MSE indicates the robustness of the model. They are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left| C_i - C_i^{GT} \right| \tag{8}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left| C_i - C_i^{GT} \right|^2} \tag{9}$$

where N is the number of the test dataset, i is the index of the image, C_i is the estimated number of the image i and C_i^{GT} is the ground truth.

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.2020.3039998, IEEE Access

C. EXPERIMENTAL RESULTS

In Table II~IV, we compare the performance of our method with several advanced works on ShanghaiTech [7], UCF_CC_50 [32] and UCSD [33]. These tables demonstrate that our method achieves the best MAE and MSE on these three benchmark datasets. Compared with the most advanced PACNN, the MAE and MSE of our FFANet on Shanghai Tech Part A decreased from 66.3 to 62.4 and RMSE from 106.4 to 102.6, which reveals that the accuracy and robustness of our method are competitive. On the challenging UCF_CC_50, our FFANet outperforms PACNN by 6.2% in MAE, which states that our method is equally effective on a dataset with a small sample size where the number of people changes dramatically. Table IV shows that our FFANet also achieved the best performance in sparse scenarios. The above results prove that our FFANet has the advantages of accuracy and robustness in both dense and sparse scenes.

TABLE II. Results on the ShanghaiTech Dataset.

Methods	ShanghaiTech _A		ShanghaiTech _B	
	MAE	MSE	MAE	MSE
Switch-CNN [13]	90.4	135.0	21.6	33.4
CP-CNN [34]	73.6	106.4	20.1	30.1
IG-CNN [35]	72.5	118.2	13.6	21.1
ic-CNN [36]	69.8	117.3	10.7	16.0
CSRNet [16]	68.2	115.0	10.6	16.0
PCC Net [15]	73.5	124.0	11.0	19.0
PACNN [17]	66.3	106.4	8.9	13.5
FFANet	**62.4**	**102.6**	**8.3**	**11.1**

TABLE III. Results on the UCF_CC_50 Dataset.

Methods	MAE	MSE
Switch-CNN [13]	318.1	439.2
CP-CNN [34]	295.8	320.9
IG-CNN [35]	291.4	349.4
ic-CNN [36]	260.9	365.5
CSRNet [16]	266.1	397.5
PCC Net [15]	240.0	315.5
PACNN [17]	241.7	320.7
FFANet	**226.8**	**316.4**

TABLE IV. Results on the UCSD Dataset.

Methods	MAE	MSE
MCNN [7]	1.60	3.31
Onoro et al. [12]	1.51	-
Switch-CNN [13]	1.62	2.10
Huang et.al [37]	1.00	1.40
CSRNet [16]	1.16	1.47
SPN [38]	1.03	1.32
FFANet	**0.97**	**1.30**

To comprehensively evaluate the performance of FFANet and other models, we further verify the results of the proposed FFANet on ShanghaiTech Part A dataset in terms of model parameters, runtime and whether to load the pre-training model. The results on other measures are shown in Table V. Compared to PACNN with the second-best MAE as shown in Table V, FFANet achieves higher accuracy with fewer parameters. However, there is still a certain gap between FFANet and the best model [15] in terms of parameters and runtime. In future work, our research direction is to use lightweight technology to reduce the amount of FFANet network parameters and maintain the inference accuracy.

TABLE V. Results on other measures on the ShanghaiTech Part A.

Methods	Parameters(M)	Runtime(ms)	Pre-train
Switch-CNN [13]	15.1	153	√
CP-CNN [34]	68.4	5113	√
CSRNet [16]	16.3	**64**	√
PCC Net [15]	**0.55**	89	×
PACNN [17]	24.1	230	√
FFANet	17.6	92	√

Fig. 5 describes the comparison of some high-quality results of FFANet on ShanghaiTech Part A with CSRNet. Compared with CSRNet, FFANet can capture the spatial distribution of the crowd and generate a clearer density map. Fig. 6 shows some high-quality results of FFANet on three datasets.

D. ABLATION STUDIES

We present the ablation studies about the hyperparameter λ, density map generation, different combinations of CBAM and network and the structure of the network.

1) ABLATION EXPERIMENTS ON LAMBDA

λ is a hyperparameter that is used to balance the MSE loss and SSIM loss. To analyze the impact of the SSIM loss function on the results, we set different λ values to observe the performance of the FFANet on ShanghaiTech Part A. Fig. 7 shows that when λ=100 the FFANet achieves the best results.

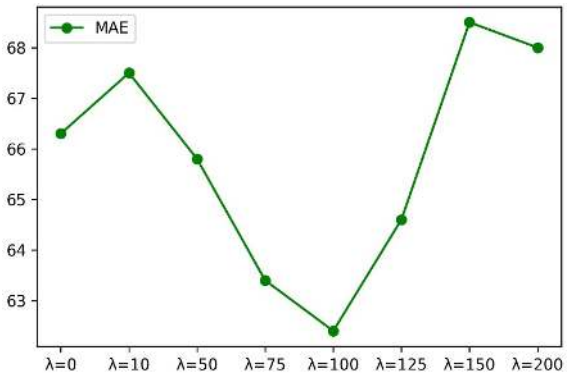

FIGURE 7. Results by varying the weight of parameter λ in the loss function.

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.2020.3039998, IEEE Access

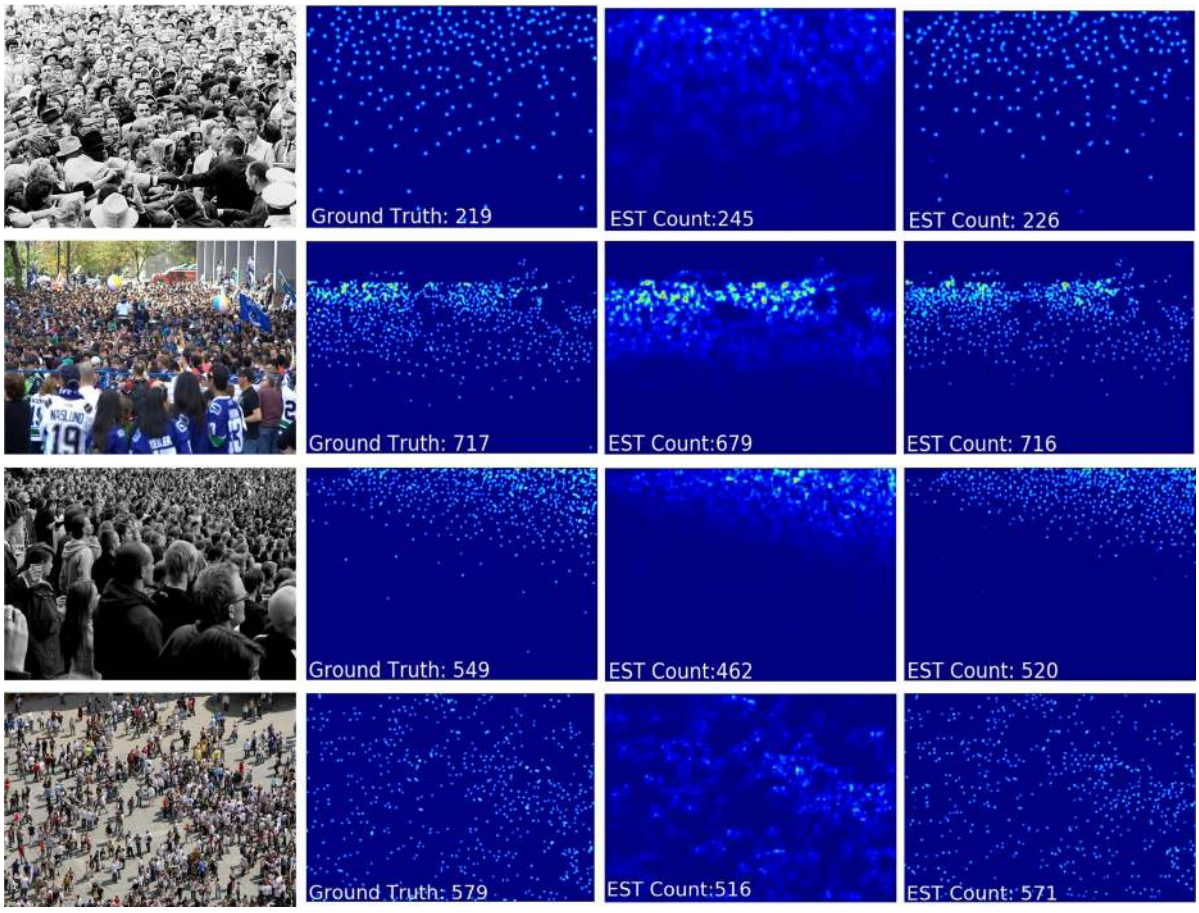

FIGURE 5. **Qualitative results on the ShanghaiTech Part A dataset. Column #1: the images from ShanghaiTech Part A; Column #2: the ground truth density maps; Column #3: estimated density maps by CSRNet; Column #4: estimated density maps by our method.**

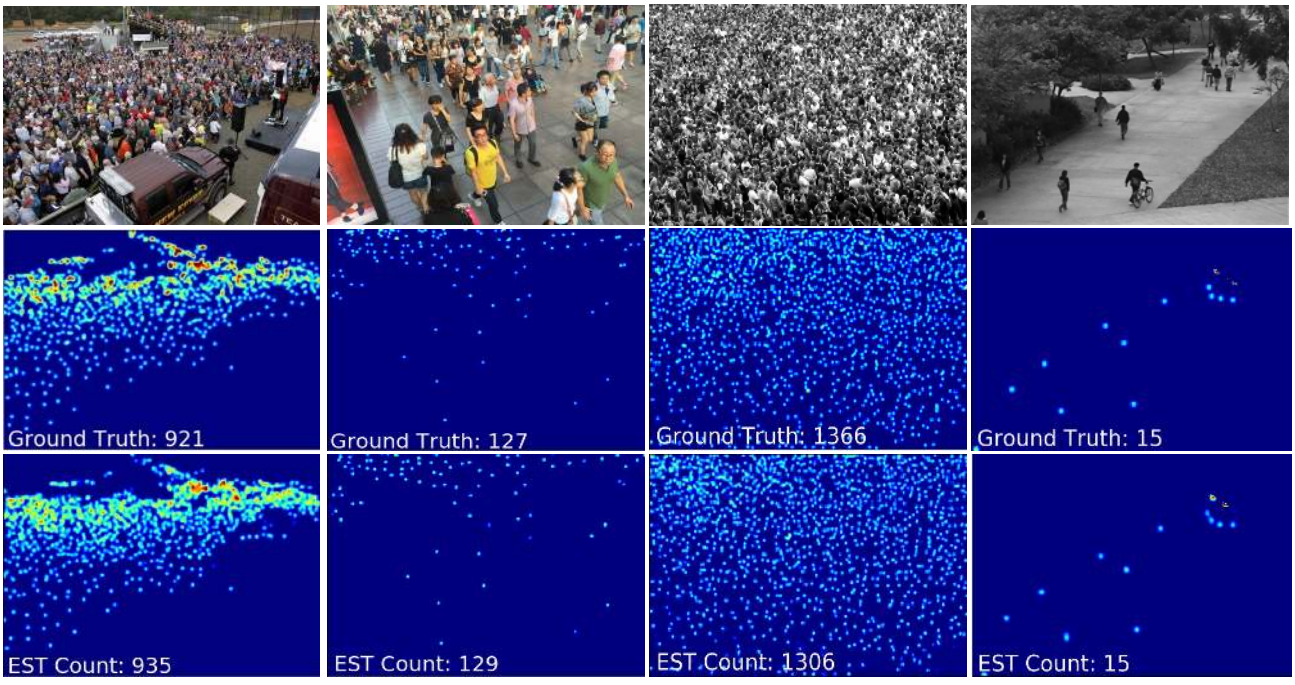

FIGURE 6. **Estimated density maps from left to right: Column #1 ShanghaiTech Part A; Column #2 ShanghaiTech Part B; Column #3 UCF_CC_50; Column #4 UCSD.**

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.2020.3039998, IEEE Access

TABLE VI. Results of density map generation on the ShanghaiTech.

Values	Part _A		Part _B	
	MAE	MSE	MAE	MSE
Adaptive-kernel [7]	66.1	110.5	8.6	11.7
Fixed-kernel (σ = 16)	68.8	106.3	9.0	11.9
Fixed-kernel (σ = 4)	**62.4**	**102.6**	**8.3**	**11.1**

TABLE VII. Results of different combinations of CBAM and network on the ShanghaiTech Part A.

Method	MAE	MSE	Parameters(M)	Runtimes(ms)
FFANet (Vanilla)	64.5	107.9	**16.9**	**86**
FFANet (Full)	63.8	109.5	17.9	101
FFANet	**62.4**	**102.6**	17.6	92

TABLE VIII. Results of different structures on the ShanghaiTech Part A.

Method	MAE	MSE	Parameters(M)	Runtimes(ms)
I. VGG16 + CM	73.6	120.9	**14.9**	**71**
II. VGG16 + stage 1 + CM (w/o FFAM)	69.1	113.8	16.2	75
III. VGG16 + stage 1, 2 + CM (w/o FFAM)	65.9	113.2	17.0	87
IV. VGG16 + stage 1 + CM (w FFAM)	66.5	105.6	16.6	76
V. FFANet (VGG16 + stage 1, 2 + CM (w FFAM))	**62.4**	**102.6**	17.6	92

2) ABLATION EXPERIMENTS ON DENSITY MAP GENERATION

In this experiment, we compare the effects of three values of Gaussian kernel which are commonly used to generate density maps in crowd counting tasks on counting results. Table VI shows that when σ = 4, the performance of the FFANet is the best.

3) ABLATION EXPERIMENTS ON CBAM ATTENTION MODULE

This part of the study is to discuss the impact of CBAM [25] on the networks. Table VII shows the performance of different combinations of CBAM and network on the ShanghaiTech Part A. FFANet (Vanilla) indicates FFANet without CBAM, FFANet (Full) means FFANet with CBAM inserted in the backbone network and FFAM. FFANet indicates that the network structure proposed in this article only inserts CBAM in FFAM. Compared with FFANet (Vanilla), the MAE and MSE of FFANet decreased from 64.5 to 62.4 and MSE from 107.9 to 102.6. Furthermore, the network parameters and runtime of FFANet increased by 0.7M and 6ms compared with FFANet (Vanilla). Compared with FFANet (Full), FFANet has achieved a comprehensive lead in all measures. After considering the performance of crowd estimation accuracy and computational cost, this paper chooses FFANet as the optimal network.

4) ABLATION EXPERIMENTS ON NETWORK STRUCTURE

We study the effects of fusing different levels of features on the accuracy of crowd counting. Table VIII represents the performance of networks with different structures on the ShanghaiTech Part A. Method I means to connect the last layer of VGG16 to the compression module to generate a crowd density map. Method II means that the features extracted from VGG16 are fused in one stage without FFAM and the fused features are connected with the compression module to generate density maps. Method III is similar to Method II, but the extracted features need two-stage feature fusion. Compared with Method I, the MAE and MSE of Method II decreased from 73.6 to 69.1 and MSE from 120.9 to 113.8, which indicates that feature fusion can effectively improve the counting accuracy. Moreover, the network parameters and runtime of method II are increased by 8.7% and 5.6% compared with Method I. Compared with Method II, the MAE and MSE of Method III reduced from 69.1 to 65.9 and MSE from 113.8 to 113.2. The results show that the performance gain is due to the increase of parameters caused by the stacking of the feature fusion stage, which strengthens the expression ability of the network. However, simply increasing the network parameters will also increase the training difficulty and weaken the robustness of the network. Compared with Method II, the network parameters and runtime of method III are greatly increased, but the MSE of Method III is only decreased by 0.6. The proposed FFAM can be a nice tradeoff between crowd estimation performance and computational cost.

To evaluate the performance of the FFAM, we added FFAM to Method II and III in the feature fusion stage, resulting in Method IV and V. Compared with Method II, the MAE and MSE of Method IV decreased from 69.1 to 66.5 and MSE from 113.8 to 105.6. Moreover, in terms of network parameters and runtime, Method IV only increases 0.6M and 1ms compared to Method I. In comparison with Method III, the MAE and MSE of Method V decreased

from 65.9 to 62.4 and MSE from 113.2 to 102.6. The decrease of MAE indicates that FFAM can solve the scale variation in an image. Meanwhile, the significant reduction in RMSE indicates that FFAM can well solve the scale changes in the dataset. The above experimental results show that FFAM is effective in dealing with scale changes in crowd counting. For the computational costs, Method V brings an increase of parameters with 0.6M and 5ms for runtime. In conclusion, FFAM can greatly improve the counting accuracy and enhance the robustness of the network under the premise of adding limited parameters.

V. CONCLUSION

In this paper, we proposed a Feature Fusion Attention Network (FFANet) to accurately estimate the number of people in the images. On one hand, the Feature Fusion Attention Module (FFAM) is proposed to realize the information enhancement of the multi-level features which are extracted by the VGG16 network. On the other hand, the enhanced features are processed by the Compression Module (CM) to generate a density map. Evaluation of the algorithm performances in comparison with other state-of-the-art methods indicates that the proposed FFANet is effective for crowd counting.

In near future, we plan to verify the adaptability of our method on other feature extractors. In addition, the performance of FFANet on UCF_CC_50 is still not perfect. This will be another research content we improve. Finally, we also plan to use model lightweight technology to reduce the time complexity of FFANet.

REFERENCES

[1] S. Zhang, G. Wu, J. P. Costeira, *et al*., "Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras", *The IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017, pp. 3667-3676.

[2] Y. Hu, H. Chang, F. Nian, *et al*., "Dense crowd counting from still images with convolutional neural networks", *Journal of Visual Communication and Image Representation*, vol. 38, pp. 530-539, Jul. 2016.

[3] H. Lu, Z. Cao, Y. Xiao, *et al*., "TasselNet: counting maize tassels in the wild via local counts regression network", *Plant methods*, vol. 13, no. 1, pp. 79, Nov. 2017.

[4] M. Hesieh, Y. Lin, W. Hsu, "Drone-based object counting by spatially regularized regional proposal network", *The IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017, pp. 4145-4153.

[5] Q. Wang, J. Gao, W. Lin, X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting", 2020, arXiv: 2001.03360. [Online]. Available: https://arxiv.org/abs/2001.03360.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, Apr. 2012.

[7] Y. Zhang, D. Zhou, S. Chen, *et al.*, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network", *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR),* Las Vegas, NV, USA, 2016, pp. 589-597.

[8] L. Yan, B. Fan, H. Liu, *et al.*, "Triplet Adversarial Domain Adaptation for Pixel-Level Classification of VHR Remote Sensing Images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3558-3573, May. 2020.

[9] H. Liu, X. Tang, S. Shen, "Depth-map Completion for Large Indoor Scene Reconstruction", *Pattern Recognition*, vol: 99, pp. 1-11, Mar. 2019, doi: 10.1016/j.patcog.2019.107112.

[10] H. Liu, Z. Xiao , B. Fan, *et al.*, "PrGCN:Probability Prediction with Graph Convolutional Network for Person Re-identification", *Neurocomputing*, early access, vol: 423, pp. 57-70, Jan. 2021.

[11] X. Cao, Z. Wang, Y. Zhao, F. Su, "Scale Aggregation Network for Accurate and Efficient Crowd Counting", *The European Conference on Computer Vision (ECCV),* Munich, Germany, 2018, pp. 734-750.

[12] D. Oñoro-Rubio, R. J. López-Sastre, "Towards Perspective-Free Object Counting with Deep Learning", *The European Conference on Computer Vision (ECCV),* Springer, Amsterdam, Netherlands, 2016, pp. 615-629.

[13] D. B. Sam, S. Surya, and R. V. Babu, "Switching Convolutional Neural Network for Crowd Counting", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 4031-4039.

[14] D. Guo, K. Li, Z. Zha, *et al.,* "DADNet: Dilated-Attention-Deformable ConvNet for Crowd Counting", *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France, 2019, pp. 1823-1832.

[15] J. Gao, Q. Wang, X. Li, "Pcc net: Perspective crowd counting via spatial convolutional network", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3486-3498, Oct. 2020.

[16] Y. Li, X. Zhang, D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Salt Lake City, UT, USA, 2018, pp. 1091-1100.

[17] M. Shi, Z. Yang, C. Xu, Q. Chen, "Revisiting Perspective Information for Efficient Crowd Counting", *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 7279-7288.

[18] N. Liu, Y. Long, C. Zou, *et al.*, "ADCrowdNet: An Attention-injective Deformable Convolutional Network for Crowd Understanding", *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 3225-3234.

[19] X. Jiang, L. Zhang, P. Lv, *et al*., "Learning multi-level density maps for crowd counting", *IEEE transactions on neural networks and learning systems*, vol. 31, no 8, pp. 2705-2715, Aug. 2020.

[20] J. U. Kim, Y. M. Ro, "Attentive Layer Separation for Object Classification and Object Localization in Object Detection", *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp. 3995-3999.

[21] M. Shaikh, V. A. Kollerathu, G. Krishnamurthi, "Recurrent Attention Mechanism Networks for Enhanced Classification of Biomedical Images", *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy, 2019, pp. 1260-1264.

[22] Z. Lan, Q. Huang, F. Chen, Y. Meng, "Aerial Image Semantic Segmentation Using Spatial and Channel Attention", *2019 IEEE 4th International Conference on Image, Vision, and Computing (ICIVC)*, Xiamen, China, 2019, pp. 316-320.

[23] H. Ling, J. Wu, L. Wu, *et al*., "Self Residual Attention Network for Deep Face Recognition," in *IEEE Access*, vol. 7, pp. 55159-55168, 2019.

[24] J. Hu, L. Shen, G. Sun, "Squeeze-and-Excitation Networks", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[25] S. Woo, J. Park, *et al*., "CBAM: Convolutional Block Attention Module", *The European Conference on Computer Vision (ECCV),* Munich, Germany, 2018, pp. 3-19.

[26] S. Woo, J. Park, *et al*., "Bam: Bottleneck Attention Module", 2018, arXiv: 1807.06514v2. [Online]. Available: http://arxiv.org/abs/1807.06514.

[27] X. Li, W. Wang, X. Hu, J. Yang, "Selective Kernel Networks", *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 510-519.

[28] Y. Cao, J. Xu, S. Lin, *et al*., "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond", *The IEEE International*

Conference on Computer Vision (ICCV), Seoul, Korea, 2019, pp. 1971-1980.

[29] X. Wang, R. Girshick, A. Gupta, K. He, "Non-Local Neural Networks", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Salt Lake City, UT, USA, 2018, pp. 7794-7803.

[30] S. Ioffe, C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", 2015, arXiv: 1502.03167. [Online]. Available: http://arxiv.org/abs/-1502.03167.

[31] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", 2014, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556.

[32] H. Idrees, I. Saleemi, *et al.*, "Multi-source Multi-scale Counting in Extremely Dense Crowd Images", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 2013, pp. 2547-2554.

[33] A. B. Chan, Z. S. J. Liang, N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking", *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-7.

[34] V. Sindagi, V. Patel, "Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs", *The IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017, pp. 1861-1870.

[35] D. B. Sam, N. N. Sajjan, "Divide and Grow: Capturing Huge Diversity in Crowd Images With Incrementally Growing CNN", *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Salt Lake City, UT, USA, 2018, pp. 3618-3626.

[36] V. Ranjan, H. Le, M. Hoai, "Iterative Crowd Counting", *The European Conference on Computer Vision (ECCV),* Munich, Germany, 2018, pp. 270-285.

[37] S. Huang, X. Li, Z. Zhang, *et al*., "Body Structure Aware Deep Crowd Counting", in *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1049-1059, Mar. 2018.

[38] X. Chen, Y. Bin, N. Sang, *et al.*, "Scale Pyramid Network for Crowd Counting", *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa Village, HI, USA, 2019, pp. 1941-1950.