Scale Attentive Aggregation Network for Crowd Counting and Localization in Smart City

WENZHE ZHAI, 1) Harbin Engineering University, Harbin, China; 2) Shandong University of Technology, Zibo, China
MINGLIANG GAO*, XIANGYU GUO, and GUOFENG ZOU, Shandong University of Technology, Zibo, China
QILEI LI, Queen Mary University of London, London, United Kingdom
GWANGGIL JEON*, 1) Shandong University of Technology, Zibo, China; 2) Incheon National University, Incheon , South Korea

Recent years have witnessed a remarkable proliferation of applications in smart cities. Crowd analysis is a crucial subject, and it incorporates two subtasks in smart city systems, *i.e.*, crowd counting and crowd localization. Nevertheless, the presence of adverse intrinsic factors, *i.e.*, scale variation and background noise severely degrades the performance of counting and localization. Although great efforts have been made on separate research on counting and localization, few works are capable of performing both tasks at the same time. To this aim, the scale attentive aggregation network (SA^{2}Net) is proposed to solve the problems of scale variation and background noise in crowd counting and localization tasks synchronously. Specifically, the SA^{2}Net has two vital modules, namely multiscale feature aggregator (MFA) module and background noise suppressor (BNS) module. The MFA module is designed in a four-pathway structure, and it aggregates the multiscale feature so as to facilitate the correlation between different scales. The BNS module utilizes the contextual information between the input keys matrix and self-attention matrix to suppress the background noise. Furthermore, a global consistency loss combined with the Euclidean loss is utilized to optimize the network in counting and localization tasks. Extensive experimental results prove that the SA^{2}Net outperforms the state-of-the-art competitors both subjectively and objectively.

CCS Concepts: • **Networks** → **Network architectures**; • **Human-centered computing** → **Visualization**.

Additional Key Words and Phrases: Smart city, Crowd counting, Crowd localization, Self-attention mechanism, Convolutional neural network.

1 INTRODUCTION

In the past decades, the emergence of Internet of Things devices and miniaturized sensing technologies have promoted the progress of smart cities [49, 50]. Crowd analysis is a hot topic and crucial task in smart city systems, *e.g.*, smart city planning, video surveillance, and public security [51, 52]. It contains many sub-tasks, *e.g.*, crowd tracking, crowd segmentation, crowd counting and crowd localization, and crowd behavior analysis [9, 25, 44].

*Corresponding Author.

Authors' addresses: Wenzhe Zhai, wenzhezhai@163.com, 1) Harbin Engineering University, Harbin, China; 2) Shandong University of Technology, Zibo, China; Mingliang Gao, mlgao@sdut.edu.cn; Xiangyu Guo, xiangyvguo@163.com; Guofeng Zou, zgf841122@163.com, Shandong University of Technology, No.266 Xincun West Road, Zhangdian District, Zibo, China; Qilei Li, qilei.li@outlook.com, Queen Mary University of London, Mile End Road, London, United Kingdom; Gwanggil Jeon, gjeon@inu.ac.kr, 1) Shandong University of Technology, Zibo, China; 2) Incheon National University, Incheon , South Korea.

ACM 1550-4859/2024/3-ART
https://doi.org/10.1145/3653454

Among them, crowd counting and localization have gained comprehensive attention in recent years, because they are applied in numerous areas, *i.e.*, public safety, congestion avoidance, and flow analysis [5, 12, 43].

The solutions to crowd counting and localization are mainly categorized as detection-based method, regression-based method, and convolutional neural network (CNN)-based method. The detection-based counting methods can be viewed as a simultaneous crowd counting and localization solution, since the bounding box is able to generate the location information of people and sum the bounding box to get the number of people. Nevertheless, there are some drawbacks to this approach, *e.g.*, the scale variation is caused by camera perspective distortion in congested scenarios to deteriorate the counting performance seriously, and the background noise can guide the model to recognize the heads incorrectly, which results in an overestimation or underestimation of the count. In order to alleviate the above defects, the regression-based method directly learns the mapping from an image to count. It turns out that this method is indeed superior to the detection-based method. Even so, they fail to produce the head location and size information, and only depending on manual features is difficult to generate a high-quality density map [5]. Thanks to the feature extraction capability of convolutional neural network (CNN), lots of scholars have proposed the density estimation methods [23, 35] to realize the counting task. For crowd localization, apart from the detection-based methods, the map-based approaches [3, 25] have become the mainstream methods. It aims to regress a high-quality density map and discover the local maxima as the head point. Therefore, generating an accurate ground truth density map is essential to precisely determine the head location.

Despite all the efforts, the performance is still unsatisfactory due to the challenging factors in crowd scenarios [4, 11, 15]. Among them, the scale variation and the background noise are the two most challenging factors that inhibit the performance. Some methods employed the multicolumn structure or dilated convolution layer to address the problem of scale variation [23, 48]. Meanwhile, many approaches adopted attention mechanisms as guidance to restrain the background noise [10, 45]. These two issues are far from resolved in the tasks of crowd counting and location. To this aim, a scale attentive aggregation network (SA^2Net) is proposed in this work. The SA^2Net has two vital modules, namely multiscale feature aggregator (MFA) module and the background noise suppressor (BNS) module. The MFA module is designed in a four-pathway structure with diverse dilated convolutional filters and aggregates multiscale information from the upper-level path to the lower-level path to enhance the scale diversity of features. The BNS module extracts the contextual information via the value of the key to instruct the dynamic attention matrix, which can effectively restrain the background noise. Meanwhile, a novel global consistency loss is proposed to allow the estimated maps to reflect the correlation of the space between pixels and the consistency between pixels. In a nutshell, the main contributions of the paper are as follows.

(1) An MFA module is built to aggregate the multiscale features so as to promote correlation at the different scales.
(2) A BNS module is established to suppress the negative effects of background noise.
(3) A global consistency loss is proposed, and it is cooperated with the Euclidean loss function to promote network convergence.
(4) A novel scale attentive aggregation network (SA^2Net) is built, and it achieves a competitive performance in tasks of crowd counting and localization.

The remainder of the paper is organized as follows. In Section 2, the related literature is reviewed. In Section 3, we detail the proposed SA^2Net model. In Section 4, comprehensive comparison and ablation studies are conducted to evaluate the proposed method. The conclusion is drawn in Section 5.

2 RELATED WORK

Recently, benefiting from the powerful feature representation ability of CNN [25, 44], CNN-based methods have widely developed and have been the mainstream in crowd counting. In this section, the two types of tasks related to the proposed SA^{2}Net are reviewed, *i.e.*, crowd counting and crowd localization.

2.1 Crowd counting

Crowd counting aims to estimate the number, density, or distribution of people in the images or videos [48]. Many approaches have been proposed to cope with the challenges of scale variation and background noise in crowd counting tasks [10, 47].

The scale variation is the limiting factor which is caused by the perspective distortion in crowd scenarios [13, 46]. Some methods attend to solve this problem. For instance, Zhang *et al.* [48] built the multicolumn structure with various convolutional filters to acquire the multiscale feature. Cao *et al.* [2] utilized four parallel convolutional filters to extract the scale features and aggregated the multiscale feature to generate high-resolution density maps. Sam *et al.* [33] employed the three CNN regressors with different column structures, which facilitate networks to pay more attention to scale information. Li *et al.* [23] used the dilated convolutional filter to enlarge the receptive fields in congested scenes, which facilitates the extraction of the detailed feature. Jiang *et al.* [20] incorporates hierarchical aggregated features into different encoding stages, which facilitates the representative capability of the network.

The background noise is similar to the pixel value of the pedestrian head area, which is not conducive to crowd counting performance. The attention mechanism can guide the network to distinguish foreground information [14, 42]. Zhai *et al.* [46] proposed dual-aware attention to suppress the background noise, which combined with channel attention and spatial attention to distinguish the foreground region. Liu *et al.* [27] built the attention map and multiscale deformable convolution to focus on the head region. Guo *et al.* [14] proposed the multi-spectral channel attention unit to identify the foreground via discrete cosine transformation (DCT) formulation. Miao *et al.* [29] build the attention model to extract the most representative feature of the crowd, which can accurately address background-pixel detection.

2.2 Crowd localization

Crowd localization is equally significant for obtaining the location of each person, rather than just inferring the total number of people in the image. It requires the location of head areas to be marked with boxes or dots, which makes the task of crowd localization even more challenging [6]. The task of crowd localization can be performed using detection-based methods and map-based approaches.

Many approaches detect the head with the points or bounding boxes to locate the head position. Liu *et al.* [26] proposed a localization branch to predict the points near annotations. An average pooling is first performed to emphasize the peak point and suppress the background noise. The non-maxima suppression strategy is then adopted to avoid points that are very close to each other. Cheng *et al.* [3] proposed a probability map that is capable of reflecting the peak probability of each head. On this basis, an algorithm named local-count-guided peak point detection is introduced, which can accurately locate the head position of each person. Sam *et al.* [32] built a top-down feature module possessing four terminals, in which the fourth terminal is responsible for distinguishing a pixel from a background or a detected head. Finally, the module outputs a localization map with bounding boxes.

More and more methods designed exclusive density maps to locate the head region. For instance, Idress *et al.* [19] discovered people through a greedy association, then matched the estimated location and the ground truth. Gao *et al.* [7] designed a binarization module to produce a threshold map, which can select the high-confidence response. The chosen responses are classified as the heads and generate an independent instance

map. Abousamra *et al.* [1] proposed a topological map to realize the localization task. Specifically, the topological map is a binary mask, which is able to connect the predicted points and the target dots. Wan *et al.* [38] proposed an unbalanced optimal transport loss to supervise the density map generation. For the localization task, a 3×3 window is used to slide on the density map without overlapping, and the point whose local maximum value is greater than a threshold value is selected as the head. Liang *et al.* [24] first applied the transformer to the crowd localization and built crowd localization transformer (CLTR) framework. The CLTR outputs a set of points and confidence scores, in which point is a pair of predicted coordinates and confidence score reflects whether a prediction point can be classified as a head.

3 FRAMEWORK OF THE PROPOSED METHOD

The framework of the proposed SA^{2}Net is shown in Fig. 1. It consists of four parts, *i.e.*, feature extractor, MFA, BNS, and generator. The feature extractor adopts the HRNet [40] to extract the basic features. The MFA module is proposed to aggregate the multiscale information, and the BNS module is built to suppress the background noise. The two modules are two critical components that are connected in parallel to improve the performance of the network. Afterward, the generator employs two transposed convolution layers to generate the estimated map. A combined loss is proposed to train the network effectively. Benefitting from the Local-Maxima-Detection-Strategy [25] and KNN strategy [30], the crowd localization maps are generated with bounding boxes.

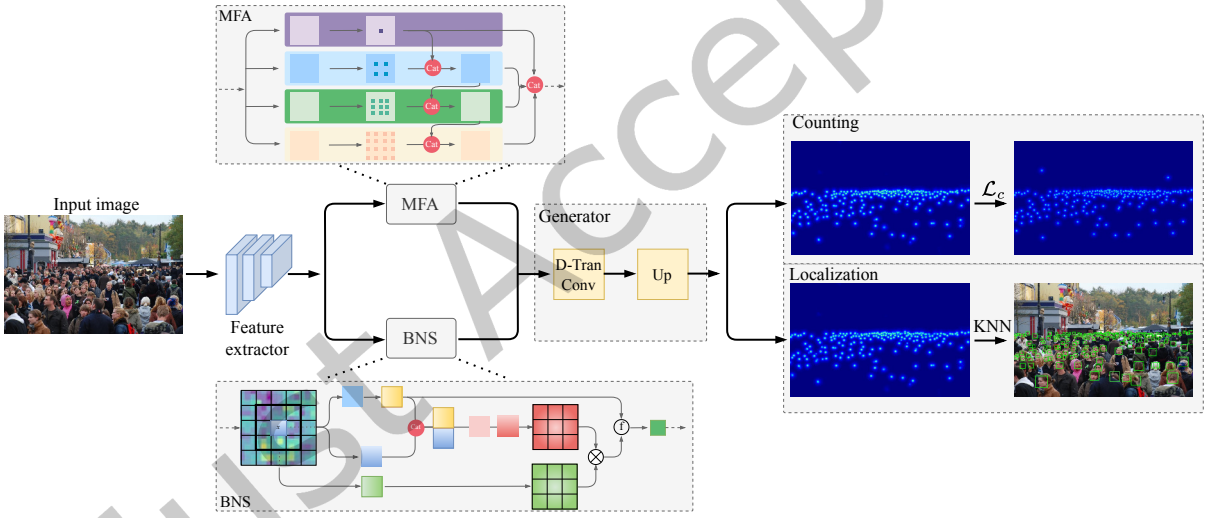

Fig. 1. Architecture of the proposed SA^{2}Net for crowd counting and localization.

3.1 Multiscale feature aggregator

Some multi-column based methods have a restriction that each column corresponds to one single scale [2]. To improve the correlation between different scales, the MFA module is proposed. The framework is shown in Fig. 2.

Given an input $x \in \mathbb{R}^{C \times H \times W}$, it is fed into the four pathways. Firstly, the number of channels is split into quarters of input feature and each feature $F_{si}(x)$ is formulated as

$$F_{si}(x) = f_{si}(x), \qquad i = 1, 2, 3, 4. \tag{1}$$

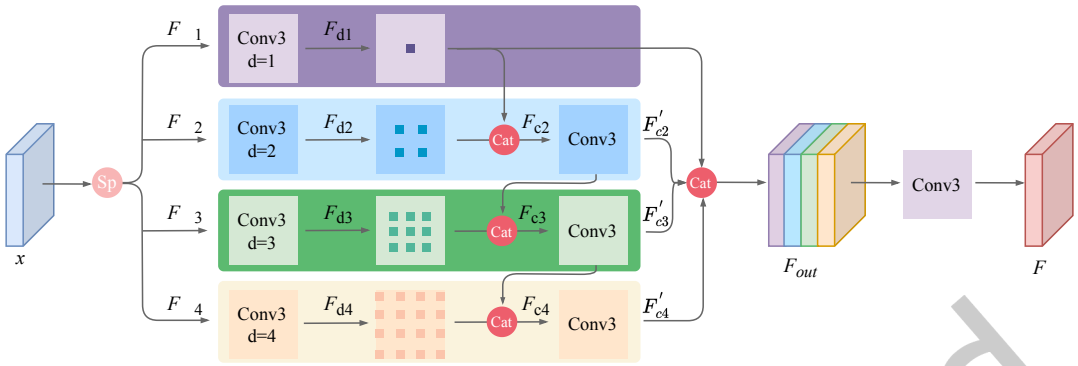

Fig. 2. Framework of the MFA module.

Where i presents i-th pathway and $f_s(\cdot)$ represents 1×1 convolution filters to reduce the channel number. The feature $F_{si} \in \mathbb{R}^{C/4 \times H \times W}$ is served as input to four pathways. For each pathway, the dilated convolution filters $f_{di}(\cdot)$ with various dilation rates of 1, 2, 3, and 4 are utilized to increase the receptive field and maintain robustness at diverse scales. It is formulated as:

$$F_{di}(x) = f_{di}(F_{si}(x)), \qquad i = 1, 2, 3, 4. \tag{2}$$

To integrate the information with diverse scales, the concatenation operation transmits scale information from the upper-level path to the lower-level path. The hierarchically aggregated scale feature information is represented as,

$$\begin{aligned} F_{c2} &= F_{d1} \oplus F_{d2}, \\ F_{c3} &= F_{d3} \oplus F_{c2}, \\ F_{c4} &= F_{d4} \oplus F_{c3}, \end{aligned} \tag{3}$$

where $\oplus$ represents the concatenation operation. Then, the 3×3 convolutional filter is performed to integrate the parallel features from the four pathways and the transformed feature of each pathway represents $F_{c2}^{'}, F_{c3}^{'}, F_{c4}^{'}$.

The aggregation feature is fused from the four pathways, and it is formulated as follows,

$$F_{out} = F_{d1} \oplus F_{c2}^{'} \oplus F_{c3}^{'} \oplus F_{c4}^{'}. \tag{4}$$

The output feature $F_{out} \in \mathbb{R}^{C \times H \times W}$ is obtained using 3×3 convolution filters to adjust the dimension.

3.2 Background noise suppressor

To suppress the adverse effects caused by background noise, we built the BNS module. It makes full use of the contextual information between input keys to instruct the learning of the dynamic attention matrix so as to enhance the foreground characterization. The BNS is depicted in Fig. 3.

First, the input feature map $x \in \mathbb{R}^{C \times H \times W}$ is transformed into three matrices. They are denoted as,

$$Q = x, K = x, V = x\mathbf{W_v}, \tag{5}$$

where Q, K, and V represent queries matrices, keys matrices, and values matrices, respectively. $\mathbf{W_v}$ denotes the parameter metric.

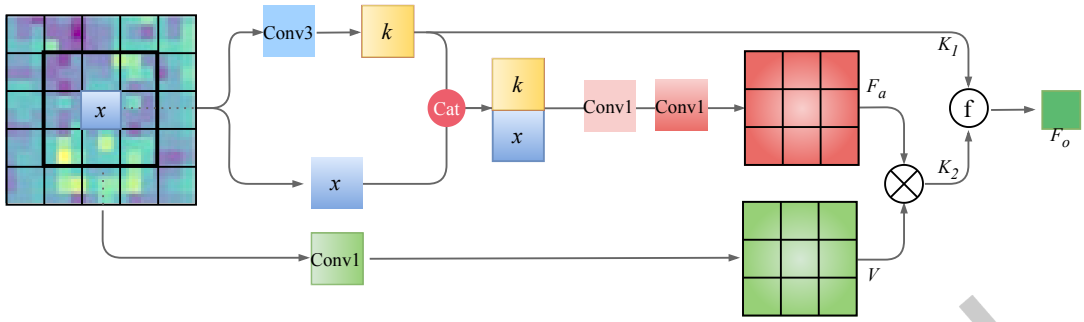

Fig. 3. Framework of the BNS module.

Afterwards, the contextual information is extracted using the 3×3 group convolution filters f_g and the learned contextualized key K_1 is formulated as follows,

$$K_1 = f_g(K). \tag{6}$$

The contextualized key K_1 and queries matrices Q are concatenated, and an attention metric is generated by two successive 1×1 convolution operations,

$$F_a = f_\gamma(f_\theta(K_1 \oplus Q)), \tag{7}$$

where $\oplus$ is a concatenation operation, f_θ represents the 1 convolutional filter with ReLU activation function and f_γ denotes the 1×1 convolutional filter without activation function, aiming to generate the attention metric. F_a reflects the self-attention metric, which learns the query feature and the context key feature to suppress background noise. The self-attention feature map K_2 is calculated by aggregating values V,

$$K_2 = V \otimes F_a, \tag{8}$$

where $\otimes$ denotes the matrix multiplication. The contextualized key K_1 and self-attention feature map K_2 are fused to produce final output F_o via the global average pooling operation.

3.3 Ground truth map

To complete the tasks of crowd counting and localization, the focal inverse distance transform (FIDT) map [25] is adopted to generate the ground-truth map, in which the head regions are acquired precisely. The FIDT map is produced as follows,

$$M(x, y) = \min_{(x', y') \in N} \sqrt{(x - x')^2 + (y - y')^2}, \tag{9}$$

$$F(x, y) = \frac{1}{M(x, y)^{(\alpha \times M(x,y) + \beta)} + C}, \tag{10}$$

where $M(x, y)$ represents the l_2 transform map, and it denotes the distance between the pixel and its nearest head position. N represents the total number of heads. α and β are hyper-parameters. $F(x, y)$ denotes the FIDT map and C is set to 1 to avoid division by 0.

3.4 Loss function

The loss function of the proposed SA^{2}Nets formulated as:

$$\mathcal{L}_c = \mathcal{L}_{mse} + \omega \mathcal{L}_{gc}, \tag{11}$$

where $\mathcal{L}_{mse}$ and $\mathcal{L}_{gc}$ represent the Euclidean loss and the global consistency loss. $\omega = 0.01$ represents the weight to balance the Euclidean loss and global consistency loss.

The Euclidean loss function is adopted to formulate the distance between the ground-truth map and the estimated map. It is denoted as,

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N}\left\|F_i^{est} - F_i^{gt}\right\|_2^2, \tag{12}$$

where the N denotes the overall headcount. F_i^{est} and F_i^{gt} represents the estimated and the ground-truth count of the i-th image. $\|\cdot\|_2^2$ represents Euclidean norm squared.

According to [2, 23], the dependence on Euclidean loss does not ensure the estimated maps reflect the spatial correlation and consistency between pixels, which are essential factors affecting the quality of density maps. To improve the global consistency in estimated maps and facilitate the regression of the head position of the crowd, the global consistency loss function $\mathcal{L}_{gc}$ is formulated as,

$$\mathcal{L}_{gc} = 1 - \frac{\sum_w^W \sum_h^H \left(F_{wh}^{est} \cdot F_{wh}^{gt}\right)}{\sqrt{\sum_w^W \sum_h^H (F_{wh}^{est})^2 \cdot \sum_w^W \sum_h^H (F_{wh}^{gt})^2}} + \left\|F_{wh}^{est} - F_{wh}^{gt}\right\|_1, \tag{13}$$

where F_{wh}^{est} and F_{wh}^{gt} denote the pixels of the estimated map and ground truth map. w and h refer to the horizontal and vertical indexes in the map, and $w \times h$ indicates the total amount of pixels. $\|\cdot\|_1$ represents the Manhattan distance.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Implementation details

The training samples are randomly cropped to 256×256 for ShanghaiTech dataset and 512×512 for others. Because the image size of ShanghaiTech dataset is smaller than other datasets. The Adam optimizer trains the network with the learning rate of 10^{-4} and the weight decay rate is 5×10^{-4}. The batch size of the training process is set to 16. The experiments are implemented based on the PyTorch framework and equipped with NVIDIA 3090Ti GPU [25].

4.2 Evaluation protocols

For the counting metrics, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are evaluated for the counting accuracy. They can reflect the accuracy and robustness of the model, respectively. The counting metrics are as follows,

$$\begin{aligned} \text{MAE} &= \frac{1}{M}\sum_{m=1}^{M}\left|C_m^{\text{est}} - C_m^{gt}\right|, \\ \text{RMSE} &= \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(C_m^{est} - C_m^{gt}\right)^2}, \end{aligned} \tag{14}$$

where M denotes the total number of images. C_m^{est} and C_m^{gt} represent the estimated result and ground truth.

For the localization metrics, the Precision, Recall, and F1-measure ($\text{F}_{1-\text{m}}$) reflect the precision of the location of the population. Specifically, Precision measures the proportion of correctly localized objects among all the predicted objects, while recall measures the proportion of correctly localized objects among all the ground-truth

objects. F_{1-m} is the harmonic mean of precision and recall, which balances the trade-off between them. The localization metrics are as follows,

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{16}$$

$$F_{1-m} = \frac{2TP}{2TP + FN + FP}, \tag{17}$$

where TP, FP, TN, and FN represent the true positive, false positive, true negative, and false negative, respectively.

4.3 Datasets

ShanghaiTech [48] includes Part A and Part B. Part A was crawled from the Internet with 300 training samples and 182 test samples which have different resolutions and relatively dense population areas. Part B was collected from the commercial streets of Shanghai with 400 training images and 316 test images. The number of people in Part A is denser than that in Part B.

UCF_CC_50 [18] consists of 50 images with a total number of 63,974 headcounts, which contains the highly-congested scenes. The five-fold cross-validation is performed to assess the performance of the proposed method. The dataset also has a large variation in image quality and resolution, which poses a challenge for feature extraction and density estimation.

UCF-QNRF [19] includes the 1,535 high-resolution samples crawled from the Internet which the crowd number ranges from 49 to 12,865 per image. It consists of 1,201 training samples and 334 test samples. It is a large-scale dataset for crowd counting that covers a wide range of crowd scenes and densities.

JHU-Crowd++ [36] has 4,372 images with a total of 1.51 million annotations, including 514 images (136,000 annotations) in bad weather such as snow, rain, and haze. It is divided into 2,722 training samples, 500 validation samples, and 1,600 test samples. The dataset also has many distractor images that contain no people or only partial people, which makes it a very challenging dataset for crowd counting.

4.4 Experimental results and analysis

4.4.1 Experiments on crowd counting. The quantitative counting results are shown in Table 1. In general, the SA^{2}Net has favorable robustness to crowd scenarios with dense and sparse data sets in ShanghaiTech dataset. On the ShanghaiTech Part A, the score of the proposed method achieves the best MAE of 58.6 and the competitive RMSE. On the ShanghaiTech Part B, the SA^{2}Net achieves the lowest MAE and RMSE. Compared with DUBNet [17] which resolves uncertainty caused by background noise, the SA^{2}Net decreases the MAE and RMSE by 3.9% and 6.4%, respectively.

The UCF-QNRF is a challenging dataset that contains a wider variety of scenarios. The SA^{2}Net scores the best MAE of 92.2 and the second-best RMSE of 169.9. It has improvements of 20.5% and 12.9% in MAE and RMSE compared with RAZ [26], which also adopted the attention mechanism to tackle the two tasks, *i.e.*, crowd counting and localization.

The UCF_CC_50 dataset includes congested scenarios which degrade the performance of the crowd counting method. The SA^{2}Net has performed more competitiveness, which achieves the best score with an MAE of 153.1 and the best score with an RMSE of 275.4. Compared with the CSRNet [23] which utilized the dilated convolution to solve the scale variation, the SA^{2}Net of performance has the improvement of 42.5% and 30.7% in MAE and RMSE.

Table 1. Objective comparison results on crowd counting. The best results are highlighted in **bold**.

Method	Part A		Part B		UCF-QNRF		UCF_CC_50		JHU-Crowd++	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN[48]	110.2	173.2	26.4	41.3	277.0	426.0	377.6	509.1	188.9	483.4
SFCN[41]	64.8	107.5	7.6	13.0	102.0	171.4	214.2	318.2	77.5	297.6
A-CCNN[21]	85.4	124.6	19.2	31.5	367.3	-	-	-	171.2	453.1
LSC-CNN[32]	66.4	117.0	8.1	12.7	120.5	218.2	225.6	302.7	112.7	454.4
CSRNet[23]	68.2	115.0	10.6	16.0	-	-	266.1	397.5	85.9	309.2
PCCNet[8]	73.5	124.0	11.0	19.0	148.7	247.3	240.0	315.5	-	-
MUD-iKNN[30]	68.0	117.7	13.4	21.4	104.0	172.0	237.7	305.7	-	-
CG-DRCN[36]	64.0	98.4	8.5	14.4	112.2	176.3	-	-	71.0	278.6
SANet[2]	67.0	104.5	8.4	13.6	-	-	258.4	334.9	91.1	320.4
MBTTBF[34]	60.2	**94.1**	8.0	15.5	97.5	**165.2**	233.1	300.9	81.8	299.1
PaDNet[37]	59.2	98.1	8.1	12.2	96.5	170.2	185.8	278.3	-	-
KDMG[39]	63.8	99.2	7.8	12.7	105.6	180.5	-	-	69.7	**268.3**
RAZ[26]	65.1	106.7	8.4	14.1	116.0	195.0	-	-	-	-
HA-CCN[35]	62.9	94.9	8.1	13.4	118.1	180.4	256.2	348.4	-	-
DUBNet[17]	64.6	106.8	7.7	12.5	105.6	180.5	243.8	329.3	-	-
CAN[28]	62.3	100.0	7.8	12.2	107.0	183.0	212.2	301.3	100.1	314.0
Ours	**58.6**	108.6	**7.4**	**11.7**	**92.2**	169.9	**153.1**	**275.4**	**66.5**	276.5

Table 2. Objective comparison results on crowd localization. The best results are highlighted in **bold**.

Method	Part A			Part B			UCF-QNRF		
	Precision(%)	Recall(%)	F_{1-m}(%)	Precision(%)	Recall(%)	F_{1-m}(%)	Precision(%)	Recall(%)	F_{1-m}(%)
MCNN[48]	-	-	-	-	-	-	59.9	63.5	61.6
TinyFaces[16]	43.1	**85.5**	57.3	64.7	79.0	71.1	36.3	77.3	49.4
LCFCN[22]	75.1	45.1	56.3	-	-	-	77.9	52.4	62.7
LSC-CNN[32]	63.9	61.0	62.4	71.7	70.6	71.2	76.6	73.5	74.0
TopoCount[1]	74.6	72.7	73.6	82.3	81.8	82.0	81.8	79.0	80.3
Ribera *et al.*[31]	67.7	44.8	53.9	-	-	-	75.5	49.8	60.1
SA^2Net (Ours)	**85.1**	82.9	**84.0**	**84.0**	**83.3**	**83.7**	**83.5**	**84.5**	**84.0**

The JHU-Crowd++ dataset has a variety of crowd scenes, which have background cluster images and bad weather scenarios. The proposed method SA^2Net outperforms all other counting approaches in MAE and achieves the second-best scores in RMSE. Particularly, compared with KDMG [39] which utilized a density map generator and density map estimator to produce density maps, the SA^2Net demonstrates the relative improvements of 4.6% in MAE. The remarkable improvement of SA^2Net implies the effectiveness of the multiscale feature aggregator and background noise suppressor.

The visualization results on crowd counting are illustrated in Fig. 4. The scenario of the ShanghaiTech Part A is challenging due to scale variation. The estimated maps of SA^2Net and headcounts perform well on congested crowd scenes with large-scale variations. The scenarios of the ShanghaiTech Part B and UCF-QNRF contain

cluttered background noise. The samples prove that the proposed method implements a satisfying result that suppresses the cluttered background effectively. The crowd distribution in the images of the UCF_CC_50 and JHU-Crowd++ is uniform, but the scale variations are large, and the crowd density is dense. The experimental results show that the density map predicted by the proposed method can reflect the population distribution.

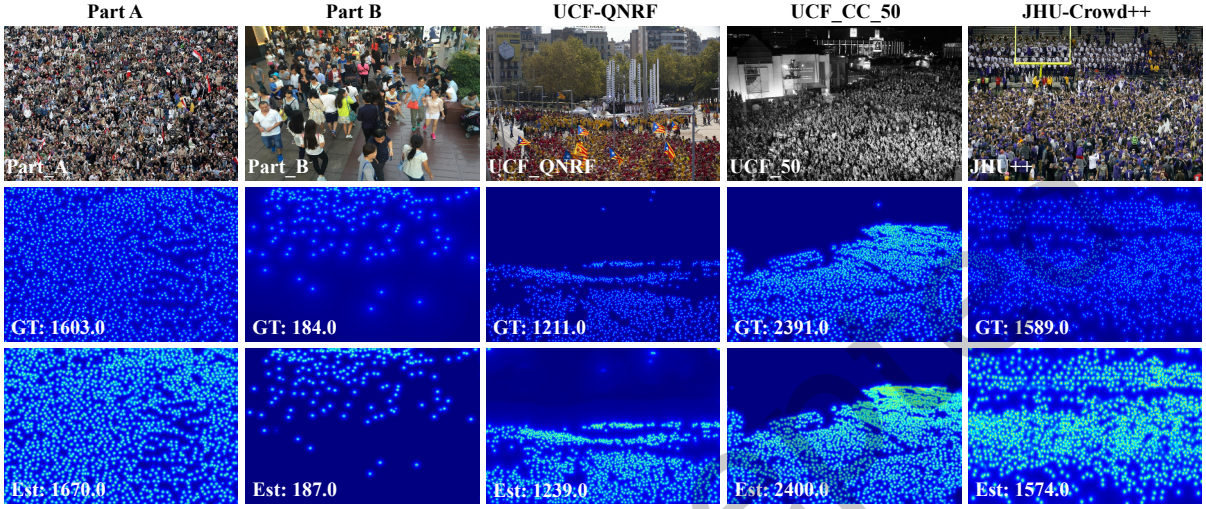

Fig. 4. Subjective results on crowd counting. The first row represents the four datasets samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. The estimated density maps and the counting number are close to the ground truth in extremely dense crowd scenarios.

4.4.2 Experiments on crowd localization. The quantitative crowd localization results are depicted in Table 2. The SA^{2}Net scores 85.1% and 84.0% in Precision and F1-measure, and achieve the second-best score of 82.9% in Recall on ShanghaiTech Part A. Compared with the TopoCount [1], the SA^{2}Net ranks the best in Precision, Recall, and F1-measure by 84.0%, 83.3%, and 83.7%, respectively. Compared with the LSC-CNN [32] which predicts the crowd localization via the bounding box in dense crowd scenes of huge diversity, the SA^{2}Net improves the Precision, Recall, and F1-measure by 17.2%, 18.0%, and 17.6%, respectively.

On the large-scale UCF-QNRF dataset, the proposed method outperforms the other competitors. Especially, it improves the performance of 2.1%, 7.0%, and 4.6% in localization metrics over the TopoCount [1] which adopts the topological constraint to predict the crowd localization.

The visualization graph on crowd localization is depicted in Fig. 5, which provides three sets of image examples. The subjective results of ShanghaiTech Part A and Part B show that the SA^{2}Net can effectively perceive head regions of different scales and improve counting accuracy. The subjective images of UCF-QNRF not only include the scale variation of pedestrians, but also more complex background regions. It can be seen that the proposed method can fully extract head regions of different scales, and effectively suppress the influence of background regions such as billboards and buildings.

4.5 Ablation studies

To evaluate the impact of the critical components and effectiveness of loss function for crowd counting and localization, ablation studies are conducted on ShanghaiTech Part A. The comparative results are depicted in Table 3. The configuration information and analysis are described as follows,

Fig. 5. Subjective results on crowd localization. The first, third, and fifth rows represent the two dataset samples with real values. The second, fourth, and sixth rows show the corresponding localization maps with estimated values. The estimated localization maps and the counting number are close to the head regions in extremely dense crowd scenarios.

- "baseline" is the model without any critical components. Also, it simply employs the Euclidean loss $\mathcal{L}_{mse}$. The counting scores of MAE and RMSE are 72.1 and 124.7, respectively. Meanwhile, the localization scores of precision, recall, and F1-m denote 76.9%, 71.8%, and 74.2%, respectively.
- "baseline+MFA" denotes that the MFA is combined with the baseline utilizing the single MSE loss. Compared with the baseline, the scores of MAE and RMSE are improved by 6.8% and 4.7%, and the scores of precision, recall, and F1-m increase by 3.0%, 4.7%, and 4.0% benefiting from the MFA module.
- "baseline+BNS" denotes that the BNS module is embedded into the baseline with single Euclidean loss $\mathcal{L}_{mse}$. The scores of MAE and RMSE are improved by 4.8% and 3.7%, which is favourable to inhibiting background clutter information to promote the robustness of the network. The scores of precision, recall, and F1-m are increased by 3.5%, 3.9%, and 3.9%. With the help of the BNS, the accuracy of localization is upgraded.
- "baseline+MFA_BNS" indicates that the BNS and MFA are embedded into the baseline by cascading operation. This type of connection provides little performance improvement in boosting the count and location. The Euclidean loss $\mathcal{L}_{mse}$ is utilized to optimize the network.
- "baseline+BNS_MFA" denotes that the BNS and MFA are embedded into the baseline by cascading operation with the Euclidean loss $\mathcal{L}_{mse}$. This type of connection has minimal effect on improving the performance of the counting and localization.
- "baseline+BNS||MFA" indicates a parallel connection type employing the Euclidean loss to train the network. It indicates that the performance outperforms the other connection types. This connection type facilitates the performance of counting and localization.
- "baseline+BNS||MFA+$\mathcal{L}_{mse}$ + $\mathcal{L}_{gc}$" represents that the MFA and BNS modules are parallelly connected. It employs the combined loss $\mathcal{L}_c$ including the Euclidean loss $\mathcal{L}_{mse}$ and the global consistency loss function $\mathcal{L}_{gc}$ to train the network which exhibits the best performance in counting metrics and localization metrics compared with all the aforementioned configurations.

Table 3. Ablation studies on the critical modules in the proposed SA^{2}Net.

Method	Counting		Localization		
	MAE	RMSE	Precision(%)	Recall(%)	F_{1-m}(%)
baseline+$\mathcal{L}_{mse}$	72.1	124.7	76.9	71.8	74.2
baseline+MFA+$\mathcal{L}_{mse}$	67.0	118.9	79.2	75.2	77.2
baseline+BNS+$\mathcal{L}_{mse}$	62.1	116.1	79.6	74.6	77.0
baseline+MFA_BNS+$\mathcal{L}_{mse}$	63.4	119.0	79.9	74.1	76.9
baseline+BNS_MFA+$\mathcal{L}_{mse}$	63.6	118.7	80.8	74.8	77.7
baseline+MFA||BNS+$\mathcal{L}_{mse}$	62.6	110.1	80.6	76.0	78.2
baseline+MFA||BNS+ $\mathcal{L}_{mse}$ + $\mathcal{L}_{gc}$	**58.6**	**108.6**	**85.1**	**82.9**	**84.0**

4.6 Failure cases

Although the proposed SA^{2}Net is capable of achieving superior counting performance against other counting methods, it still has some failure cases. Some failure cases are visualized in Fig. 6. One can see that there is a large gap between the estimated value and the ground truth, especially for crowd counting and localization in low-light scenarios. Crowd counting in low-light environments is challenging, as the features of the head region are much closer to those of the background region than in low-light environments. The density maps estimated by the algorithms are contaminated with some unnecessary background interference when the crowd scenes

are acquired in low-light. Therefore, we consider that there is still a large room for crowd counting in low-light scenes.

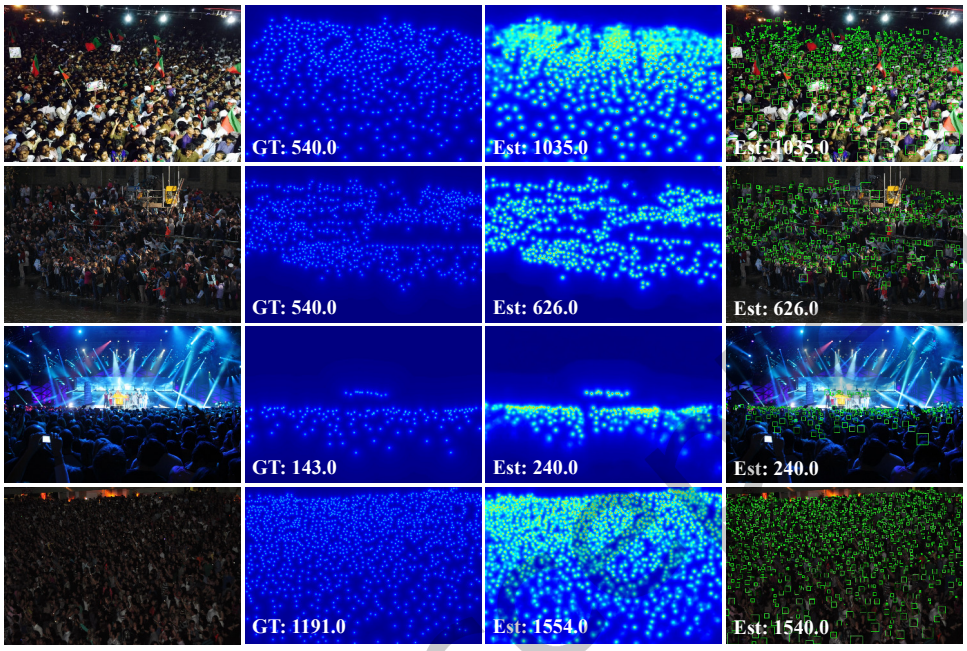

Fig. 6. The failure cases. The first column, the second column, the third column, and the fourth column depict the exemplars, the ground truth, the estimated counting results, and the localization results, respectively.

5 CONCLUSION

In this paper, we present the SA^{2}Net to tackle the problems of scale variation and background noise in crowd counting and localization tasks. The proposed SA^{2}Net is characterized by two critical components, *i.e.*, MFA module and BNS module. The MFA module is employed to aggregate multiscale features to promote correlation between different scales via a four-pathway structure. The BNS module is built to improve foreground characterization by instructing the relationship between the input keys and self-attention matrix for suppressing the background noise. Meanwhile, a global consistency loss combined with the Education loss facilitates the SA^{2}Net to collaborate on the task of crowd counting and localization. Extensive experiments and ablation experiments have proved the superiority of the SA^{2}Net in both crowd counting and localization. In the future, more attention and effort are expected to be devoted to research on handling low-light counting problems, as the head region has less detailed information and is highly susceptible to background interference in low-light.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Nos.61601266 and 61801272) and the National Natural Science Foundation of Shandong Province (Nos.ZR2021QD041 and ZR2020MF127).

REFERENCES

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. 2021. Localization in the Crowd with Topological Constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 872–881. doi: https://doi.org/10.1609/aaai.v35i2.16170.

[2] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and F. Su. 2018. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 734–750. doi:https://doi.org/10.1007/978-3-030-01228-1_45.

[3] Jian Cheng, Haipeng Xiong, Zhiguo Cao, and Hao Lu. 2021. Decoupled Two-Stage Crowd Counting and Beyond. *IEEE Transactions on Image Processing* 30 (2021), 2862–2875. doi:https://doi.org/10.1109/TIP.2021.3055631.

[4] Zizhu Fan, Hong Zhang, Zheng Zhang, Guangming Lu, Yudong Zhang, and Yaowei Wang. 2022. A survey of crowd counting and density estimation based on convolutional neural network. *Neurocomputing* 472 (2022), 224–251. doi: https://doi.org/10.1016/j.neucom.2021.02.103.

[5] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. 2020. CNN-based Density Estimation and Crowd Counting: A Survey. *ArXiv* abs/2003.12783 (2020).

[6] Junyu Gao, Maoguo Gong, and Xuelong Li. 2022. Congested crowd instance localization with dilated convolutional swin transformer. *Neurocomputing* 513 (2022), 94–103. doi:https://doi.org/10.1016/j.neucom.2022.09.113.

[7] Junyu Gao, Tao Han, Yuan Yuan, and Qi Wang. 2020. Learning Independent Instance Maps for Crowd Localization. *ArXiv* abs/2012.04164 (2020).

[8] Junyu Gao, Q. Wang, and Xuelong Li. 2020. PCC Net: Perspective Crowd Counting via Spatial Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology* 30 (2020), 3486–3498. doi:https://doi.org/10.1109/TCSVT.2019.2919139.

[9] Mingliang Gao, Alireza Souri, Mayram Zaker, Wenzhe Zhai, Xiangyu Guo, and Qilei Li. 2023. A comprehensive analysis for crowd counting methodologies and algorithms in Internet of Things. *Cluster Computing* (2023), 1–15. doi:https://doi.org/10.1007/s10586-023-03987-y.

[10] Meng-Hao Guo, Tianhan Xu, Jiangjiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph Robert Martin, Ming-Ming Cheng, and Shimin Hu. 2022. Attention Mechanisms in Computer Vision: A Survey. *ArXiv* abs/2111.07624 (2022).

[11] Xiangyu Guo, Mingliang Gao, Wenzhe Zhai, Qilei Li, and Gwanggil Jeon. 2023. Scale Region Recognition Network for Object Counting in Intelligent Transportation System. *IEEE Transactions on Intelligent Transportation Systems* (2023). doi:https://doi.org/10.1109/TITS.2023.3296571.

[12] Xiangyu Guo, Mingliang Gao, Wenzhe Zhai, Qilei Li, Kyu Hyung Kim, and Gwanggil Jeon. 2023. Dense Attention Fusion Network for Object Counting in IoT System. *Mobile Networks and Applications* (2023), 1–10. doi:https://doi.org/10.1007/s11036-023-02090-1.

[13] Xiangyu Guo, Mingliang Gao, Wenzhe Zhai, Qilei Li, Jinfeng Pan, and Guofeng Zou. 2022. Multiscale aggregation network via smooth inverse map for crowd counting. *Multimedia Tools and Applications* (2022), 1–15. doi: https://doi.org/10.1007/s11042-022-13664-8.

[14] Xiangyu Guo, Mingliang Gao, Wenzheng Zhai, Jianrun Shang, and Qilei Li. 2022. Spatial-Frequency Attention Network for Crowd Counting. *Big data* 10, 5 (2022), 453–465. doi: https://doi.org/10.1089/big.2022.0039.

[15] Xiangyu Guo, Kai Song, Mingliang Gao, Wenzhe Zhai, Qilei Li, and Gwanggil Jeon. 2023. Crowd counting in smart city via lightweight Ghost Attention Pyramid Network. *Future Generation Computer Systems* 147 (2023), 328–338. doi:https://doi.org/10.1016/j.future.2023.05.013.

[16] Peiyun Hu and Deva Ramanan. 2017. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 951–959. doi: https://doi.org/10.1109/CVPR.2017.166.

[17] Min hwan Oh, P. Olsen, and K. Ramamurthy. 2020. Crowd Counting with Decomposed Uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 11799–11806. doi:https://doi.org/10.1609/AAAI.V34I07.6852.

[18] H. Idrees, Imran Saleemi, C. Seibert, and M. Shah. 2013. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2547–2554. doi:https://doi.org/10.1109/CVPR.2013.329.

[19] H. Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah. 2018. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 532–546. doi:https://doi.org/10.1007/978-3-030-01216-8_33.

[20] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David S. Doermann, and Ling Shao. 2019. Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 6126–6135. doi:https://doi.org/10.1109/CVPR.2019.00629.

[21] Saeed Amirgholipour Kasmani, Xiangjian He, W. Jia, Dadong Wang, and Michelle Zeibots. 2018. A-CCNN: Adaptive CCNN for Density Estimation and Crowd Counting. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 948–952. doi:https://doi.org/10.1109/ICIP.2018.8451399.

[22] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. 2018. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 547–562. doi: https://doi.org/10.1007/978-3-030-01216-8_34.

[23] Yuhong Li, Xiaofan Zhang, and D. Chen. 2018. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1091–1100. doi:https://doi.org/10.1109/CVPR.2018.00120.

[24] Dingkang Liang, Wei Xu, and Xiang Bai. 2022. An End-to-End Transformer Model for Crowd Localization. *Proceedings of the European Conference on Computer Vision (ECCV)* abs/2202.13065 (2022).

[25] Dingkang Liang, Wei Xu, Yingying Zhu, and Yu Zhou. 2022. Focal Inverse Distance Transform Maps for Crowd Localization. *IEEE Transactions on Multimedia* (2022), 1–13. doi:https://doi.org/10.1109/TMM.2022.3203870.

[26] Chenchen Liu, Xinyu Weng, and Yadong Mu. 2019. Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization. *CVPR* (2019), 1217–1226. doi: https://doi.org/10.1109/CVPR.2019.00131.

[27] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. 2019. ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3220–3229. doi:https://doi.org/10.1109/CVPR.2019.00334.

[28] Weizhe Liu, M. Salzmann, and P. Fua. 2019. Context-Aware Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5094–5103. doi:https://doi.org/10.1109/CVPR.2019.00524.

[29] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. 2020. Shallow Feature Based Dense Attention Network for Crowd Counting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 11765–11772. doi:https://doi.org/10.1609/AAAI.V34I07.6848.

[30] Greg Olmschenk, Hao Tang, and Zhigang Zhu. 2020. Improving Dense Crowd Counting Convolutional Neural Networks using Inverse k-Nearest Neighbor Maps and Multiscale Upsampling. In *15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Vol. 5. doi: https://doi.org/10.5220/0009156201850195.

[31] Javier Ribera, David Guera, Yuhao Chen, and Edward J Delp. 2019. Locating objects without bounding boxes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6479–6489. doi: https://doi.org/10.1109/CVPR.2019.00664.

[32] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R. Venkatesh Babu. 2021. Locate, Size, and Count: Accurately Resolving People in Dense Crowds via Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), 2739–2751. doi:https://doi.org/10.1109/tpami.2020.2974830.

[33] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. 2017. Switching Convolutional Neural Network for Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4031–4039. doi:https://doi.org/10.1109/CVPR.2017.429.

[34] Vishwanath A. Sindagi and Vishal M. Patel. 2019. Multi-Level Bottom-Top and Top-Bottom Feature Fusion for Crowd Counting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1002–1012. doi:https://doi.org/10.1109/ICCV.2019.00109.

[35] Vishwanath A. Sindagi and Vishal M. Patel. 2020. HA-CCN: Hierarchical Attention-Based Crowd Counting Network. *IEEE Transactions on Image Processing* 29 (2020), 323–335. doi: https://doi.org/10.1109/TIP.2019.2928634.

[36] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. 2022. JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5 (2022), 2594–2609. https://doi.org/10.1109/TPAMI.2020.3035969

[37] Yukun Tian, Yiming Lei, Junping Zhang, and James Z. Wang. 2020. PaDNet: Pan-Density Crowd Counting. *IEEE Transactions on Image Processing* 29 (2020), 2714–2727.

[38] Jia Wan, Ziquan Liu, and Antoni B. Chan. 2021. A Generalized Loss Function for Crowd Counting and Localization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 1974–1983. doi: https://doi.org/10.1109/CVPR46437.2021.00201.

[39] Jia Wan, Qingzhong Wang, and Antoni B. Chan. 2020. Kernel-based Density Map Generation for Dense Object Counting. *IEEE transactions on pattern analysis and machine intelligence* (2020), 1–1. doi:https://doi.org/10.1109/TPAMI.2020.3022878.

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3349–3364. doi: https://doi.org/10.1109/TPAMI.2020.2983686.

[41] Q. Wang, Junyu Gao, Wei Lin, and Y. Yuan. 2019. Learning From Synthetic Data for Crowd Counting in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8190–8199. doi:https://doi.org/10.1109/CVPR.2019.00839.

[42] Wenzhe Zhai, Mingliang Gao, Marco Anisetti, Qilei Li, Seunggil Jeon, and Jinfeng Pan. 2022. Group-split attention network for crowd counting. *Journal of Electronic Imaging* 31, 4 (2022), 041214. doi:https://doi.org/10.1117/1.JEI.31.4.041214.

[43] Wenzhe Zhai, Mingliang Gao, Xiangyu Guo, and Qilei Li. 2023. Scale-Context Perceptive Network for Crowd Counting and Localization in Smart City System. *IEEE Internet of Things Journal* (2023). doi:https://doi.org/10.1109/JIOT.2023.3268226.

[44] Wenzhe Zhai, Mingliang Gao, Qilei Li, Gwanggil Jeon, and Marco Anisetti. 2023. FPANet: feature pyramid attention network for crowd counting. *Applied Intelligence* (2023), 1–18. doi:https://doi.org/10.1007/s10489-023-04499-3.

[45] Wenzhe Zhai, Mingliang Gao, Alireza Souri, Qilei Li, Xiangyu Guo, Jianrun Shang, and Guofeng Zou. 2022. An attentive hierarchy ConvNet for crowd counting in smart city. *Cluster Computing* (2022), 1–13. doi:https://doi.org/10.1007/s10586-022-03749-2.

[46] Wenzhe Zhai, Qilei Li, Ying Zhou, Xuesong Li, Jinfeng Pan, Guofeng Zou, and Mingliang Gao. 2022. DA2Net: a dual attention-aware network for robust crowd counting. *Multimedia Systems* (2022). https://doi.org/10.1007/s00530-021-00877-4 doi:https://doi.org/10.1007/s00530-021-00877-4.

[47] Wenzhe Zhai, Jinfeng Pan, Qilei Li, Guofeng Zou, Liju Yin, and Mingliang Gao. 2021. A Channel-aware Attention Network for Crowd Counting. In *2021 China Automation Congress (CAC)*. 4048–4052. doi:https://doi.org/10.1109/CAC53003.2021.9728649.

[48] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 589–597. doi:https://doi.org/10.1109/CVPR.2016.70.

[49] Xiaokang Zhou, Wei Liang, Jinhua She, Zheng Yan, I Kevin, and Kai Wang. 2021. Two-layer federated learning with heterogeneous model aggregation for 6g supported internet of vehicles. *IEEE Transactions on Vehicular Technology* 70, 6 (2021), 5308–5317. doi:https://doi.org/10.1109/TVT.2021.3077893.

[50] Xiaokang Zhou, Wei Liang, Kevin I-Kai Wang, Zheng Yan, Laurence T. Yang, Wei Wei, Jianhua Ma, and Qun Jin. 2023. Decentralized P2P Federated Learning for Privacy-Preserving and Resilient Mobile Robotic Systems. *IEEE Wireless Communications* 30, 2 (2023), 82–89. doi:https://doi.org/10.1109/MWC.004.2200381.

[51] Xiaokang Zhou, Xiaozhou Ye, I Kevin, Kai Wang, Wei Liang, Nirmal Kumar C Nair, Shohei Shimizu, Zheng Yan, and Qun Jin. 2023. Hierarchical federated learning with social context clustering-based participant selection for internet of medical things applications. *IEEE Transactions on Computational Social Systems* (2023). doi:https://doi.org/10.1109/TCSS.2023.3259431.

[52] Xiaokang Zhou, Xuzhe Zheng, Xuesong Cui, Jiashuai Shi, Wei Liang, Zheng Yan, Laurance T Yang, Shohei Shimizu, I Kevin, and Kai Wang. 2023. Digital Twin Enhanced Federated Reinforcement Learning with Lightweight Knowledge Distillation in Mobile Networks. *IEEE Journal on Selected Areas in Communications* (2023). doi:https://doi.org/10.1109/JSAC.2023.3310046.