Applied Intelligence (2023) 53:19199–19216
https://doi.org/10.1007/s10489-023-04499-3

ORIGINAL SUBMISSION

FPANet: feature pyramid attention network for crowd counting

Wenzhe Zhai[1] · Mingliang Gao[1] · Qilei Li[2] · Gwanggil Jeon[3] · Marco Anisetti[4]

Accepted: 31 January 2023 / Published online: 27 February 2023

Abstract

Crowd counting in congested scenarios is an essential yet challenging task in detecting abnormal crowd for contemporary urban planning. The counting accuracy has been significantly improved with the rapid development of deep learning over the last decades. However, current models are fragile in the real-world application mainly due to two inherent weaknesses: (1) Scale variations always exert negative influences on counting accuracy. (2) Overwhelming amount of parameters in the deep neural network will lead to low efficiency. To address these two limitations, in this paper, we propose a Feature Pyramid Attention Network (FPANet). Specifically, the FPANet consists of three modules, namely the feature pyramid module, attention module, and multiscale aggregation module. The feature pyramid module is built in a lightweight architecture to extract multiscale features. The attention module focuses on the crowd region and suppresses misleading information. The multiscale aggregation module is derived to adaptively fuse the discriminative knowledge extracted in different granularities. Additionaly, the efficiency of FPANet is boosted by the multi-group structure. Experimental results on five crowd benchmark datasets, i.e., ShanghaiTech, UCF_CC_50, UCF-QNRF, WorldExpo'10, and NWPU-Crowd, and two cross-domain datasets, i.e., CARPK, and PUCPR+, demonstrate that the FPANet achieves superior performances in terms of accuracy, efficiency and generalization.

Keywords Crowd counting · Lightweight network · Attention mechanism · Pyramid attention · Scale variation

1 Introduction

Crowd counting has been an emerging topic in the computer vision community in recent years. It plays crucial roles in a wide range of applications, e.g., crowd simulation, crowd dynamics modeling, vehicle detection, and vehicle counting [51, 52, 69]. Inspired by the remarkable performance of convolutional neural networks (CNNs) [23, 29], researchers have lately developed numerous CNN-based crowd counting methods [13, 20, 64]. The key idea of the CNN-based method is to employ a CNN to regress the density map and then integrate the pixels on the map from which the final count value is derived [20, 69] Some CNN-based crowd counting methods [66, 69] attempted to solve the problem of large-scale variations by taking advantage of multiscale architectures. These methods designed subnetworks to learn an informative feature cross-multiscale intermediate to get rid of the effect of the distribution gap. Other approaches adopted attention mechanisms as a guide in improving the prediction accuracy [4, 70].

Although counting accuracy has been advanced by these recent attempts, they are still fragile when being adopted in real-world applications, reflected by poor adaptation performance in counting accuracy. We attribute this mainly to the large-scale variation which is intrinsically challenging for model learning. Some examples of crowd scenarios with large-scale variations are depicted in Fig. 1, which can diminish the quality of estimated density maps, leading to erroneous estimation in backgrounds.

Moreover, contemporary counting models typically seek trivial performance improvement by deepening or widening the network, which is computationally heavy and inefficient to be deployed in low-cost devices, such as mobile phones

✉ Mingliang Gao
mlgao@sdut.edu.cn

1 School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China

2 School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

3 Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, South Korea

4 Department of Computer Science, Universitá degli Studi di Milano, Milano, 20133, Italy

Fig. 1 Crowd scenarios with large-scale variations. The top row shows some typical samples, and the bottom row represents the ground truth density maps

and embedded equipment. To solve the problem, many light-weight models were proposed [2, 5]. However, these light-weight methods are struggle in finding a balance to the computationnal cost or the counting performance and being incapable to depict the crowd distribution [54].

In this work, we consider these two issues jointly and proposed a unified framework named Feature Pyramid Attention Network (FPANet) to perform precious counting in a lightweight design. Specifically, the proposed FPANet model consists of three modules, i.e., feature pyramid module, attention module, and multiscale aggregation module. The feature pyramid module adopts a multi-column grouping structure, which extracts multi-scale features and promotes the network to be more efficient. The attention module focuses simultaneously on the spatial dependencies in the whole feature map in order to acquire precise head regions by the pyramid spatial attention (PSA) unit, and assists in processing the relationships between channel maps and highlights the discriminatory information in a particular channel by light-weight channel attention (LCA) unit. The multiscale aggregation module integrates the multiscale spatial information and the cross-channel knowledge into the block for each feature group. To summarize, the contributions of this paper are as follows.

1. We deal with the scale variations in complex crowd scenes by building a feature pyramid (FP) module and multiscale aggregation (MA) module to exploit the robust scale features without increasing parameters and introduce the scale communication architecture between multiscale inputs.
2. We build a dual-attention module, named a pyramid spatial attention (PSA) unit and a light-weight channel attention (LCA) unit, to accurately locate dense regions of the input without relying on prior knowledge.
3. Experimental results on five benchmark crowd datasets, i.e., ShanghaiTech, UCF_CC_50, UCF-QNRF, WorldExpo'10, NWPU-Crowd, and two cross-domain datasets, i.e., CARPK, and PUCPR+, prove that the FPANet achieves remarkable performances in accuracy, efficiency, and generalization.

The remaining of the paper is structured as follows: An overview of the related works is introduced in Section 2. The proposed FPANet is shown in Section 3. The implementation details are introduced in Section 4. The experimental results and analysis are presented in Section 5. This work is concluded in Section 6.

2 Related work

Recently, the CNN-based methods have been the mainstream in this domain thanks to the powerful feature representation ability of CNN [15, 25, 53]. In this section, we review three kinds of CNN-based methods which are closely related to the proposed FPANet, i.e., multiscale CNNs methods, attention CNNs methods and light-weight CNNs methods.

2.1 Multiscale CNNs methods

The multiscale CNNs methods usually adopt multi-column network (MCNN) to capture multiscale information so as to deal with the problem of scale variation. The multiscale CNNs architecture for crowd counting was first proposed in [69], in which a three-column CNNs with different receptive fields was built. Subsequently, Sam et al. [41] improved the MCNN by adding the recurrent networks to fuse features from multi-column CNN and leveraged a switch classifier to tackle large-scale variations. Meanwhile, to improve the MCNN, Marsden et al. [33] utilized a multiscale average mechanism to address the problems of scale and perspective variation. Unlike multi-column networks, Zhang et al. [67]

proposed the scale adaptive convolutional neural network (SaCNN) by progressively increasing the number of layers from lower to higher levels with different scales. Although the multiscale CNNs methods work well in dealing with the scale variations, they ignore the rich context and location information [4, 64].

2.2 Attention CNNs methods

The attention mechanism is to adjust the weight according to the importance of features, and it is widely applied in many practical domains [21, 24, 26]. Recently, the attention mechanism is also generalized to the crowd counting domain. Liu et al. [27] pioneered the use of an attention module called QualityNet to perceive changes in crowd density. To improve the counting capability, Gao et al. [4] built the spatial-/channel-wise attention regression networks (SCAR), in which a spatial attention module is introduced to extract global context information and a channel attention module is built to alleviate the side effect of background interference. Similarly, Zhai et al. [65] proposed the DA^{2}Net which consists of a spatial attention module locating the heads and a channel attention module highlighting the discriminative region which further increases in counting accuracy. Different from the SCAR and DA^{2}Net which utilize the channel and spatial attention, Jiang et al. [16] presented an attention scaling-based counting network that exploits attention masks to label the regions with different density levels.

2.3 Light-weight CNNs methods

Although the multiscale CNNs methods and attention CNNs methods have achieved remarkable progress, their performance comes with the cost of burdensome computation. In this regard, how to reduce the burden of network computing draws growing interest. For instance, Cao et al. [2] proposed an efficient counting framework called scale aggregation network. It introduced a patch-based scheme to simplify the training process, and a lightweight loss to reduce the computing burden. Besides, Wang et al. [54] designed an efficient encoder-decoder architecture with limited computation resources. The aforementioned approaches simply focus on the simplification of the network and have not taken into account the scale information of the head region. To address this issue, Gao et al. [5] proposed a model to extract the large-range and perspective information with fewer parameters. Meanwhile, Zhai et al. [63] proposed the group-spilt attention network to process the scale features of each group in parallel with few computational costs. The above methods reduced the computational burden at the expense of accuracy.

In this paper, the proposed FPANet explicitly integrates the feature pyramid module, attention module, and multiscale aggregation module to promote the performance of crowd counting in accuracy, efficiency and generalization. Besides, we conduct comprehensive studies on the architecture variations to verify the superiority of the proposed FPANet.

3 Proposed method

3.1 Overview

The structure of the proposed FPANet is depicted in Fig. 2. It is composed of three modules, i.e., feature pyramid (FP) module, attention module, and multiscale aggregation (MA) module. First, the ResNet-50 [8] is tailored, and the first three layers are retained as the backbone network. Next, the FP module is built to extract multiscale features. Then, a dual-aware attention module consisting of a pyramid

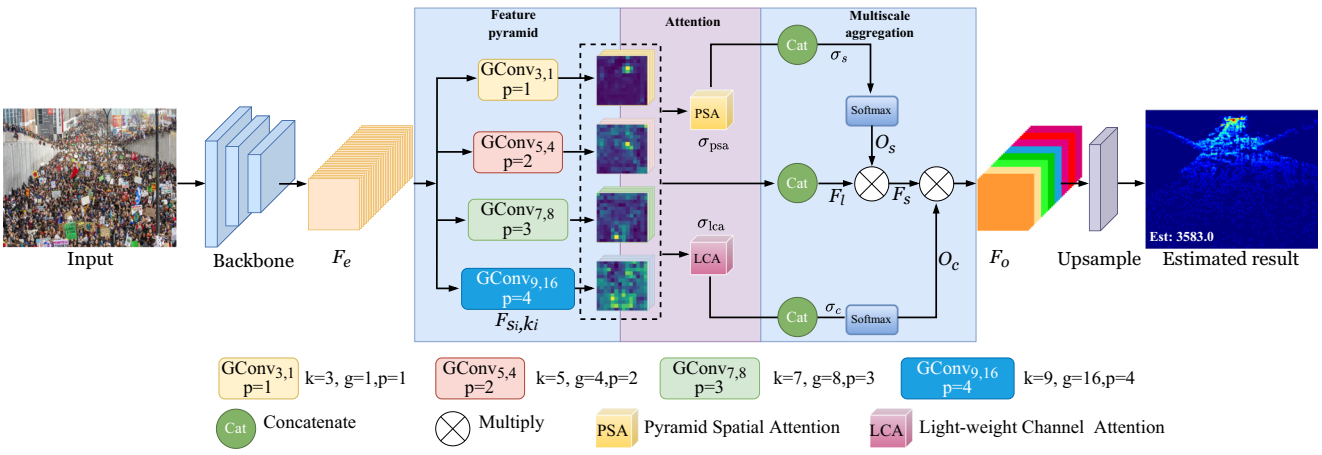


Fig. 2 Diagram of the proposed FPANet for crowd counting

spatial attention (PSA) unit and a light-weight channel attention (LCA) unit is set up in each branch. After that, the MA module integrates the multiscale spatial information and the cross-channel attention into the block of each feature group. Finally, an upsample operation is adopted to output the predicted density map.

3.2 Feature pyramid module

Given an image I, the original crowd image is fed into the ResNet-50 [10, 56] with the first three layers to extract low-level features $F_e \in \mathbb{R}^{C\times H\times W}$ from the input images. Afterwards, the feature pyramid (FP) module employs the multi-group structure to extract the multiscale feature in which the channel dimension of each subgroup is C. Meanwhile, the group convolution operation is utilized to reduce the number of calculations.The pyramid feature map F_{s_i,k_i} is generated as,

$$F_{s_i,k_i} = f_{\text{Gconv}}[F_e(I)], g = 2^{\frac{k-1}{2}}, \qquad i = 1, 2, \cdots, N, \tag{1}$$

where g is the number of group convolution. The k represents the convolutional kernel size of 3×3, 5×5, 7×7, and 9×9. Especially, the default value of g is set to 1 when the value of k equals 3. $f_{\text{Gconv}}(\cdot)$ represents the group convolution operation with different kernel sizes to aggregate cross-channel information from the input feature map $F_e(I)$. It reduces the number of parameters to be optimized in the network. As a consequence, the model can be trained at a low cost. In addition, to maintain the same size of the input and output feature maps, the values of padding are set to 1,2,3,4 and the stride is set to 1, respectively. The $F_{s_i,k_i} \in \mathbb{R}^{C/4\times H\times W}$ represents the global hierarchical features for the next operations. s_i and k_i represent the group size and the convolution kernel size of the i-th scale level.

3.3 Attention module

The attention module is designed in a dual-aware pattern, which is composed of a pyramid spatial attention (PSA) unit and a light-weight channel attention (LCA) unit. The PSA unit generates a spatial attention map by utilizing the inter-spatial features. Meanwhile, the LCA unit produces a channel attention map by exploiting the interrelationship of channels. The attention weights at various scales are gained by extracting the channel attention weight information from the multiscale preprocessed feature maps. Mathematically, the attention weight is denoted as,

$$\sigma_i = f_{\text{att}}(F_{s_i,k_i}), \tag{2}$$

where $\sigma_i \in \mathbb{R}^{C/4\times 1\times 1}$ is the attention weight, which represents the general term for PSA and LCA. $f_{\text{att}}(\cdot)$ indicates an operation to obtain multiscale attention weight. It is implemented through a series of convolutions followed by Rectified Linear Units (ReLU) [35] and Batch Normalization (BN) [14].

Pyramid spatial attention unit The feature map $F_{s_i,k_i} \in \mathbb{R}^{C/4\times H\times W}$ is fed into the PSA unit. An attention mechanism based CNN model learns attention weights from the input i-th scale sizes and multiplies each channel in F_{s_i,k_i} to produce a multiscale attention map. The 1-dimension attention map P_s is formulated as,

$$P_s = f_r(f_p(F_{s_i,k_i}, 4)) \oplus f_r(f_p(F_{s_i,k_i}, 2)) \oplus f_r(f_p(F_{s_i,k_i}, 1)), \tag{3}$$

where $f_p(\cdot)$ denotes the adaptive average pooling with three scales, i.e., 4×4, 2×2, and 1×1. $f_r(\cdot)$ resizes a tensor size. $\oplus$ denotes the element-wise concatenation. The architecture of PSA unit is illustrated in Fig. 3. It consists of three types of average pooling operations. The average pooling of 4×4 is utilized to acquire more feature

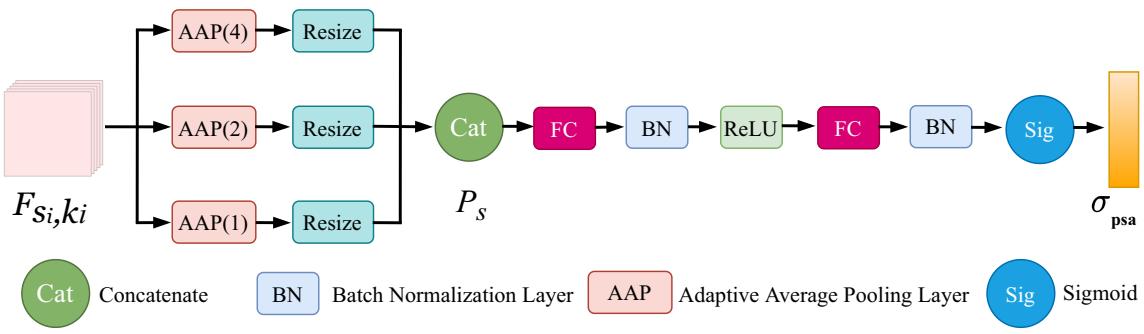


Fig. 3 Framework of the pyramid spatial attention (PSA) unit

representation and structural details. The average pooling of 2×2 devotes to a trade-off between structural information and structural regularization. The average pooling 1×1 is the general global average pooling with a strong structural regularization. The three outputs are resized to three 1-dimension maps. Then, the three branches are concatenated to generate the feature map P_s.

However, the concatenation operation is limited by learning channel dependencies and affecting the effectiveness of the attention mechanism [7]. To tackle this problem, we encode the P_s using the incentive block and construct a 1-dimensional attention map σ_{psa}. The incentive block utilizes two fully connected layers and a sigmoid layer for regularizing the output to the range (0, 1). The PSA unit is formulated as follows,

$$\sigma_{psa} = \text{Sigmoid}(f_{c2}\rho(f_{c1}(P_s))), \tag{4}$$

where f_{c1} and f_{c2} represent two fully-connected layer with BN layers, respectively. ρ denotes a rectified linear unit (ReLU) function. σ_{psa} generates the spatial attention weight.

Light-weight channel attention unit The aforementioned PSA unit seeks to encode the correlations in spatial dimension so that the head region can be precisely detected. Nevertheless, it could lead to the incorrect estimation of the background because of similarities between the foreground and background area textures. In addition, to reduce the parameters, some attention models, such as SENet [10] and CBAM [61], involve dimensionality reduction to reduce the model complexity. However, it is inefficient and unnecessary to capture the dependencies across all the channels [56].

To address these issues, we employ the complementary LCA units inspired by the strategy in [56] to ensure accuracy and efficiency. The architecture of LCA unit is shown in Fig. 4. The formulation is shown as follows,

$$\begin{aligned} g_c &= \frac{1}{WH}\sum_{i=1,j=1}^{W,H} F_{s_i,k_i}, \\ \sigma_{lca} &= \text{Sigmoid}(\text{C1D}_k(g_c)), \end{aligned} \tag{5}$$

where F_{s_i,k_i} represents the feature map corresponding to each channel. g_c indicates the channel-aware global average pooling operation that acquires the aggregated features of the background region. The LCA module generates channel weight σ_{lca} by performing a fast 1-dimension convolution (C1D) which utilizes the convolution kernel size ($k = 3$) with fewer parameters.

3.4 Multiscale aggregation module

The multiscale aggregation (MA) module aggregates and facilitates feature fusion of different channel dimensions. It contains three operations, namely channel concatenation, feature recalibration and feature map re-weighting operation.

The channel concatenation operation fuses the multiscale pre-processed feature map of different dimensions. It represents the general term for multiscale layer aggregation and attention map aggregation. The multiscale layer aggregation is formulated as,

$$F_l = F_{s_1,k_1} \oplus F_{s_2,k_2} \oplus \cdots \oplus F_{s_N,k_N}, \tag{6}$$

where $F_l \in \mathbb{R}^{C\times H\times W}$ employs multiscale feature map and $\oplus$ is the concatenation operator. In this way, the multiscale layer aggregation can integrate multiscale cross-hierarchy contextual features at different scales and generate superior pixel-level attention for high-level feature maps.

The attention map aggregation operation aggregates the PSA unit and the LCA unit, respectively. The PSA attention map aggregation is applied to gain the spatial attention weight from the input feature map with different scales. The whole multiscale PSA attention vector σ_s is attained in a concatenation manner as,

$$\sigma_s = \sigma_{\text{psa}_1} \oplus \sigma_{\text{psa}_2} \oplus \cdots \oplus \sigma_{\text{psa}_N}, \tag{7}$$

where σ_{psa_i} is the i-th multiscale PSA attention weight vector.

In the same way, the LCA attention map aggregation is used to obtain the channel attention weight from the input feature map with different scales. The multiscale LCA attention weight vector σ_c is formulated as follows,

$$\sigma_c = \sigma_{\text{lca}_1} \oplus \sigma_{\text{lca}_2} \oplus \cdots \oplus \sigma_{\text{lca}_N}. \tag{8}$$

The feature recalibration operation adaptively selects different spatial and channel attention scales through a cross-channel strategy, which is guided by the attention

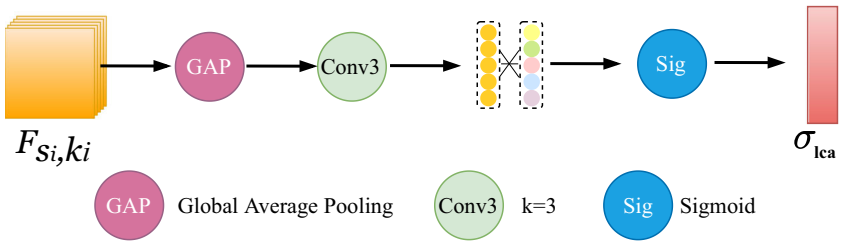


Fig. 4 Framework of the light-weight channel attention (LCA) unit

weight vectors σ_s and σ_c. The recalibrated channel spatial attention O_s is formulated as,

$$O_s = \text{Softmax}(\sigma_s) = \frac{\exp(\sigma_s)}{\sum_{i=1}^{n} \exp(\sigma_s)}, \tag{9}$$

where the Softmax$(\cdot)$ is utilized to acquire the recalibrated multiscale PSA weight O_s. It contains the location information on the attention weight in spatial. The PSA unit generates the spatial attention map $O_s \in \mathbb{R}^{H \times W}$ to reflect the importance in the spatial space of the concerned areas (e.g., head).

Similarly, the channel attention of feature recalibration is fused and spliced in a concatenation manner. The recalibrated channel attention vector O_c is formulated as,

$$O_c = \text{Softmax}(\sigma_c) = \frac{\exp(\sigma_c)}{\sum_{i=1}^{n} \exp(\sigma_c)}, \tag{10}$$

where the Softmax$(\cdot)$ is utilized to obtain the recalibrated multiscale LCA weight O_c.

The feature map re-weighting operation is designed to redress the weight of the recalibrated PSA and LCA weights. Specifically, the re-assigned spatial feature map F_s is obtained by the multiplication of spatial attention assignment weight O_s and the multiscale feature map F_c, as

$$F_s = F_l \otimes O_s, \tag{11}$$

where $\otimes$ represents the element-by-element multiplication. The re-assigned spatial feature map F_s acts as a bridge between the PSA unit and the LCA unit.

The re-assigned channel feature map can be formulated by the multiplication of the re-assigned spatial feature map F_s and the recalibrated multiscale channel weight as,

$$F_o = F_s \otimes O_c, \tag{12}$$

where $F_o \in \mathbb{R}^{C \times H \times W}$ represents the enhanced feature map. It provides more discriminative information than the previous feature map F_e in both spatial and channel dimensions.

4 Implementation details

4.1 Ground truth density map

Similar to [20, 69], the geometry-adaptive kernel is employed to generate the density map. It is formulated as follows,

$$H(z) = \sum_{i=1}^{N} \delta(z - z_i) * \text{G}_\sigma(z), \tag{13}$$

where z_i denotes the i-th annotated head location. The delta function $\delta(z - z_i)$ and normalized Gaussian kernel G with the fixed parameter σ are convolved to generate ground truth density maps.

4.2 Loss function

Following the earlier works [28, 69], the Euclidean loss is employed to measure the estimation error between the estimated density map and the ground truth. It is formulated as,

$$\text{loss} = \frac{1}{M} \sum_{i=1}^{M} \| F(I_i) - Y_i \|_2^2, \tag{14}$$

where M denotes the batch size. $F_\theta(I_i)$ represents the predicted density map. Y_i represents the corresponding density map of ground truth.

4.3 Data augmentation

The accuracy of population counting is plagued by scale problems, so we randomly change the scale to help crowd estimation. Particularly, we utilize random horizontal flips to increase data diversity. The original images are resized to 576×768 along with generating the final density images of the same size. Images are normalized with the mean and deviation using the PyTorch framework [4].

4.4 Training details

The default batch size is set to 4. The first three layers of a pre-trained ResNet-50 [8] are adopted as the backbone to extract the low-level features. Adam [18] is adopted as the optimizer. The models are trained for 400 epochs. and the learning rate is initially set to 10^{-5} and multiplied by 0.995 per epoch. All the experiments are conducted on two NVIDIA RTX3090Ti GPUs. The source code will be available at https://github.com/wzzhai/FPANet-for-Crowd-Counting.

5 Experimental results and analysis

In this section, we first compare the proposed FPANet with the state-of-the-art methods on five benchmark datasets to verify the accuracy. Besides, we apply FPANet on two cross-datasets to validate the generalization ability. Meanwhile, we carry out measure the parameters and complexity of the calculations with other mainstreams to verify the efficiency of the proposed FPANet. The typical scenarios from the crowd datasets are illustrated in Fig. 5. Finally, the ablation experiments are performed to verify the effectiveness of the various modules in FPANet.

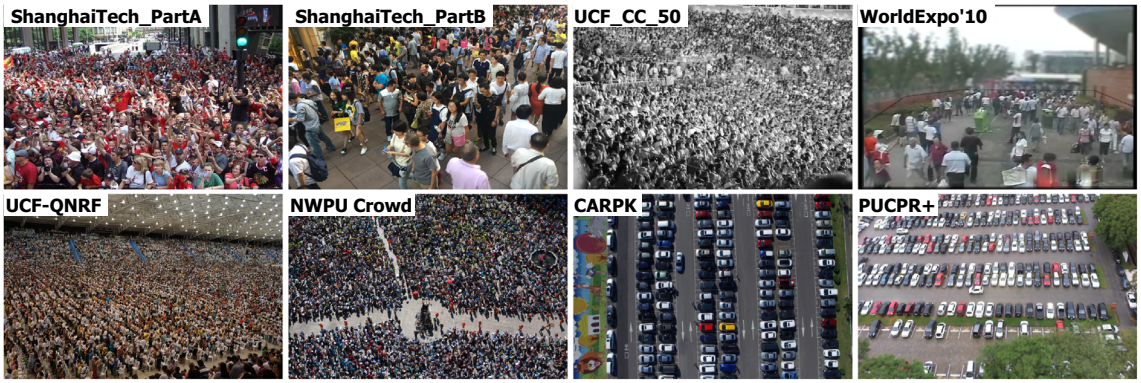


Fig. 5 Typical scenarios in the crowd benchmark datasets (ShanghaiTech, UCF_CC_50, WorldExpo'10, UCF-QNRF, and NWPU-Crowd) and the cross-domain car benchmark datasets (CARPK, and PUCPR+ datasets)

5.1 Evaluation metrics

Following the earlier works [20, 44, 66], the count performance is measured in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which are formulated as,

$$\text{MAE} = \frac{1}{N}\sum \left| y_i - \hat{y}_i \right|, \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum \left| y_i - \hat{y}_i \right|^2}, \tag{16}$$

where N denotes the total number of test images. y_i represents the ground truth count and $\hat{y}_i$ is the estimated value for the i-th image.

5.2 Performance on ShanghaiTech dataset

The ShanghaiTech dataset [69] is composed of 1198 images with 330,165 annotated heads. It is composed of two subsets, namely Part_A and Part_B. In Part_A, 482 images (300 training images and 182 test images) are selected from the internet. In Part_B, 716 images (400 training images and 316 test images) are collected from the downtown region in Shanghai. The comparative results are presented in Table 1. On ShanghaiTech Part_A, the FPANet scores 70.9 in MAE, ranking first place among the competitors. Meanwhile, it achieves a score of 120.6 in RMSE, which ranks second place among the competitors. On ShanghaiTech Part_B, the proposed method performs best in both MAE and RMSE, and outperforms others competitors by a large margin. Compared with SaCNN [67] which also combines multiple layers to solve the scale problem, the FPANet reduces the MAE and RMSE by 18.33% and 13.4%. Figure 6 illustrates the visualization results on the ShanghaiTech dataset.

5.3 Performance on UCF_CC_50 dataset

The UCF_CC_50 dataset [12] merely contains 50 images with extremely high crowd density. Despite the limited data samples, it still provides a variety of scenarios, i.e., concerts, stadiums, etc. The head annotations of each image range from 94 to 4,543, and there are an average of 1,280 pedestrians per image. Following the principle [12], we randomly split the dataset into five pieces for cross-validation. We compared the proposed FPANet with the SOTA methods in terms of MAE and RMSE. As depicted in Table 2, the FPANet scores 159.5 and 218.4 in MAE and RMSE, both outperforming the competitors. Specially,

Table 1 Experimental results on the ShanghaiTech dataset (The best results are marked in **bold**)

Method	Part_A		Part_B	
	MAE	RMSE	MAE	RMSE
Zhang et al. [66]	181.8	277.7	32.0	49.8
Marsden et al. [33]	126.5	173.5	23.8	33.1
MCNN [69]	110.2	173.2	26.4	41.3
CMTL [44]	101.3	152.4	20.0	31.1
NLT [58]	93.8	157.2	11.8	19.2
TDF-CNN [40]	97.5	145.1	20.7	32.8
Switching-CNN [41]	90.4	135.0	21.1	30.1
BSAD [11]	90.4	135.0	20.2	35.6
DecideNet [27]	–	–	20.8	29.4
BSAD [11]	90.4	135.0	20.2	35.6
TDF-CNN [40]	97.5	145.1	20.7	32.8
SaCNN [67]	86.8	139.2	20.7	32.8
A-CCNN [17]	85.4	124.6	11.0	19.0
MATT [19]	80.1	129.4	11.7	17.5
PCC-Net [5]	73.5	124.0	19.2	31.5
DNCL [68]	73.5	**112.3**	18.7	26.0
FPANet(ours)	**70.9**	120.6	**8.8**	**15.5**

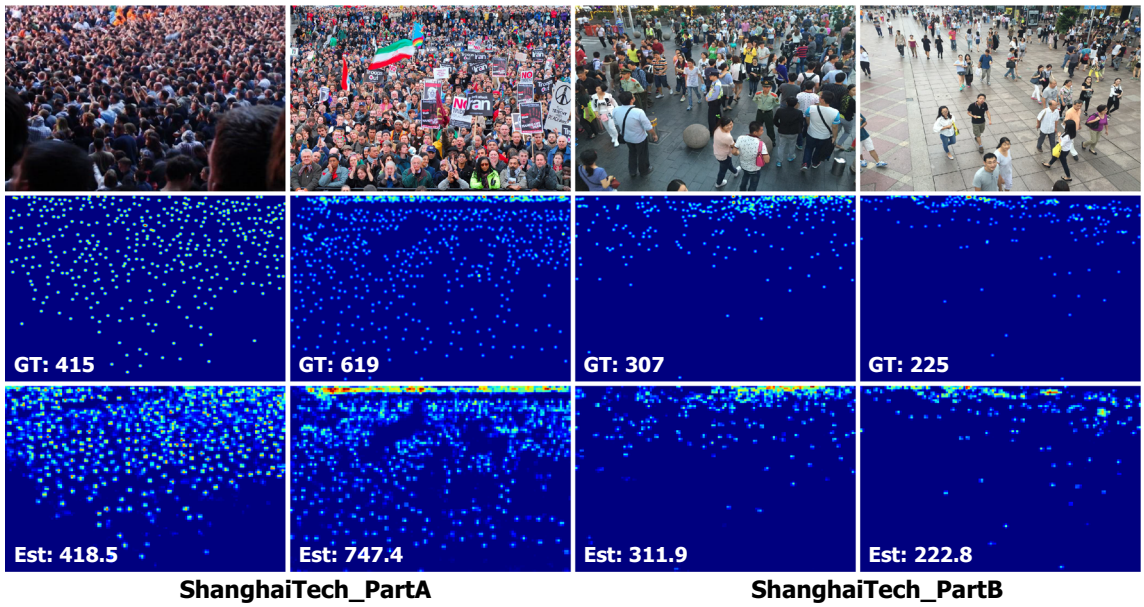


Fig. 6 Visualization results on the ShanghaiTech Part_A (first two columns) and Part_B (first two columns) datasets. The first row represents the four samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. The sample images depict crowd scenarios with uneven distribution and varying head sizes. The proposed method can provide accurate crowd density maps and counting results in both dense and sparse scenes

compared with SCAR [4], ASNet [16] and DA^{2}Net [65] which also utilize the attention mechanism, the proposed proposed FPANet reduces the MAE by 37.2%, 6.92% and 5.9%, and RMSE by 39.6%, 10.3% and 7.8%, respectively. The visualized results of the FPANet on UCF_CC_50 dataset are depicted in Fig. 7. It proves that the proposed method implemens a satisfying result in highly dense crowd scenes.

5.4 Performance on WorldExpo'10 dataset

The WorldExpo'10 dataset [66] is a cross-scene benchmark dataset captured by 108 surveillance cameras from Shanghai WorldExpo. It contains 3,980 frames in 103 scenarios in the training set, and 600 frames from the rest 5 scenarios in the testing set. The ROI areas are defined according to the general criteria [58]. The experimental results of 5 scenes on the WorldExpo'10 dataset are depicted in Table 3. It depicts that the FPANet sores the first place in Scenes 1, 2, and 3, except for Scenes 4 and 5. Simultaneously, it obtains the best scores in terms of average MAE, which is reduced by 1.8% compared to the suboptimal method DA^{2}Net[65]. The exemplar qualitative results on the WorldExpo'10 dataset are presented in Fig. 8.

5.5 Performance on UCF-QNRF dataset

The UCF-QNRF dataset [13] is a challenging dataset due to its diversified viewpoints, various crowd scales and densities. It contains 1,535 high-resolution images with an average size of 2013×2902. It involves more outdoor real scenes (e.g., trees, sky, buildings and roads) compared to other datasets, which makes it more realistic. The comparative results with other SOTA methods are

Table 2 Experimental results on the UCF_CC_50 dataset (The best results are marked in **bold**)

Methods	MAE	RMSE
Idrees et al. [12]	419.5	541.6
Zhang et al. [66]	467.0	498.5
MCNN [69]	377.6	509.1
MATT [19]	355.0	550.2
CMTL [44]	322.8	397.9
SaCNN [67]	314.9	424.8
CP-CNN [45]	295.8	320.9
ACM-CNN [70]	291.6	337.0
DNCL [68]	288.4	407.7
DADNet [6]	285.5	389.7
MobileCount [54]	283.1	382.6
CSRNet [20]	266.1	397.5
ic-CNN [37]	260.9	365.5
SCAR [4]	259.0	374.0
PCCNet [5]	240.0	315.5
ASNet [16]	174.8	251.6
DSNet [3]	183.3	240.6
DA^{2}Net [65]	169.5	237.0
FPANet (ours)	**159.5**	**218.4**

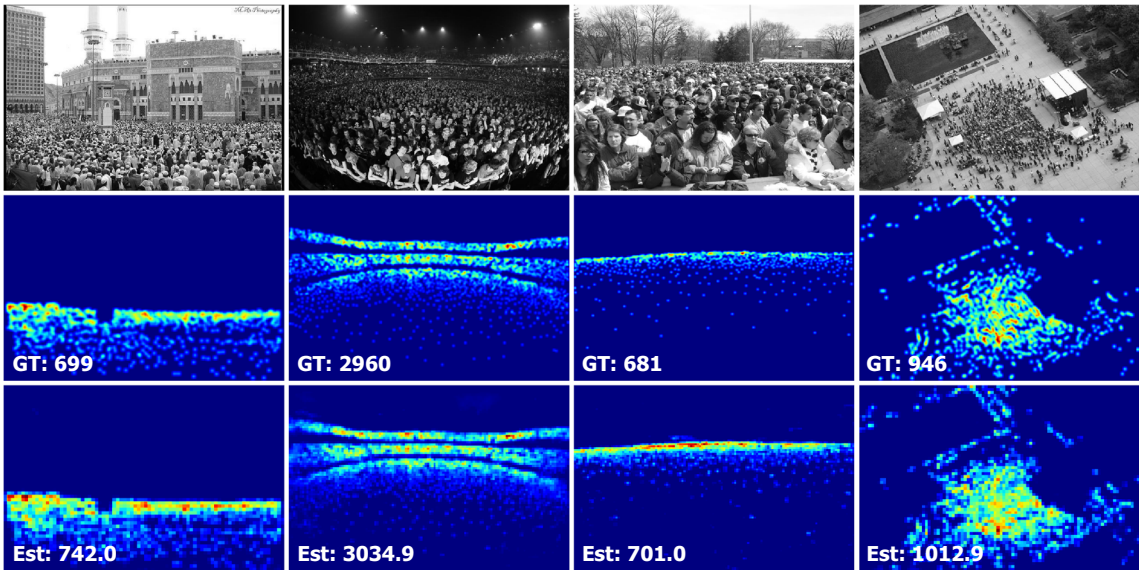


Fig. 7 Visualization results on UCF_CC_50 dataset. The first row represents the four samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. The estimated density maps and the counting number are close to the ground truth in extremely dense crowd scenario

demonstrated in Table 4. The proposed FPANet achieves the best result with an MAE of 108.9 and the third-best result with an RMSE of 197.6, which are comparative to the best RMSE from the DUBNet [36]. Particularly, it reduces the MAE and RMSE by 2.5% and 3.3% compared with DA^{2}Net [65], which also employs the channel and spatial attention mechanism. The visual density maps are depicted in Fig. 9.

5.6 Performance on NWPU-Crowd dataset

The NWPU-Crowd dataset [57] is the largest benchmark dataset, and it has a total number of 5,109 images with 2,133,375 head annotations. It is more challenging due to the influence of negative samples, and large appearance changes. Compared with the aforementioned datasets, the difference is mainly reflected in two aspects. For one

Table 3 Experimental results on the WorldExpo'10 dataset (The best results are marked in **bold**)

Methods	S1	S2	S3	S4	S5	MAE(Avg.)
Zhang et al. [66]	9.8	14.1	14.3	22.4	3.7	12.9
MCNN [69]	3.4	20.6	12.9	13.0	8.1	11.6
MSCNN [59]	7.8	15.4	14.9	11.8	5.8	11.7
SCAR [4]	1.9	13.8	9.6	29.8	3.9	11.8
BSAD [11]	1.4	21.7	11.9	11.0	3.5	10.5
ic-CNN [37]	17.0	12.3	9.2	8.1	4.7	10.3
PCC Net [5]	1.9	18.3	10.5	13.4	3.4	9.5
DNCL [68]	1.9	12.1	20.7	8.3	2.6	9.1
CSRNet [20]	2.9	11.5	8.6	16.6	3.4	8.6
ACM-CNN [70]	2.4	10.4	11.4	15.6	3.0	8.56
SaCNN [67]	2.6	13.5	10.6	12.5	3.3	8.5
SANet [2]	2.6	13.2	9.0	13.3	3.0	8.2
LSC-CNN [42]	2.9	11.3	9.4	12.3	4.3	8.0
M-SFANet [47]	1.88	13.24	10.07	**7.5**	3.87	7.32
STDNet [31]	1.83	12.78	10.3	7.88	2.5	7.05
DA^{2}Net [65]	1.45	11.8	7.95	11.6	**2.35**	7.03
FPANet (ours)	**1.2**	**10.5**	**8.5**	12.0	2.5	**6.9**

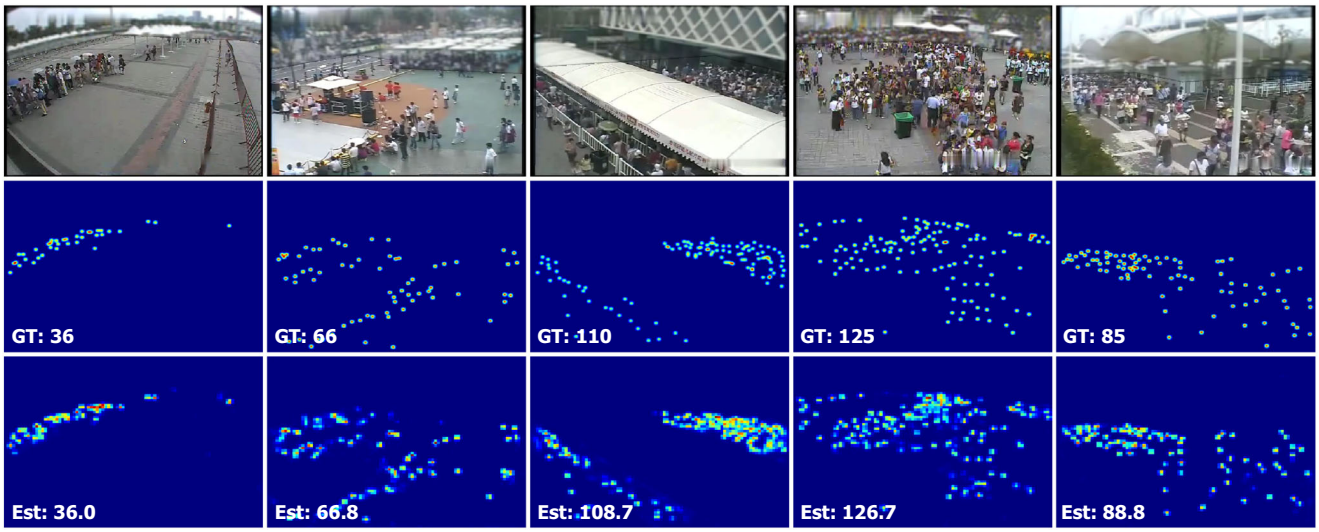


Fig. 8 Visualizations on the WorldExpo'10 dataset. The first row represents the four samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. It shows that the estimated maps and counting numbers are very close to the ground truth in uneven distributed crowd scenes

thing, it has much more diversity in scales, density and background. For another, it includes 351 negative samples (namely nobody scenes), which increases the variety of datasets. As illustrated in Table 5, the proposed method achieves the best results in both MAE and RMSE, with a reduction of 3.4% and 1.5% compared with the suboptimal MAE of KDMG [50] and the suboptimal RMSE of DA^{2}Net [65]. Figure 10 displays some qualitative validation set results of the proposed method.

Table 4 Experimental results on the UCF-QNRF dataset (The best results are marked in **bold**)

Methods	MAE	RMSE
Zhang et al. [66]	467.0	498.5
MCNN [69]	277.0	509.1
SCAR [4]	264.8	418.3
CMTL [44]	252.0	514.0
Switching-CNN [41]	228.0	445.0
PCCNet [5]	148.7	247.3
NLT [58]	172.3	263.1
ZoomCount [39]	130.0	204.0
CRSNet [20]	129.0	209.0
DENet [28]	121.0	205.0
LSC-CNN [42]	120.5	218.2
MobileCount [54]	117.7	207.6
DUBNet [36]	116.0	**178.0**
DADNet [6]	113.2	189.4
DA^{2}Net [65]	111.7	204.3
FPANet (ours)	**108.9**	197.6

5.7 Cross-dataset analysis

To verify the generalization capability of the proposed FPANet, cross-dataset evaluation is performed. Following the work in [43], we firstly adopt the ShanghaiTech Part_A as the training set, while the ShanghaiTech Part_B and UCF-QNRF datasets as the test sets, respectively. Then, the ShanghaiTech Part_B and UCF-QNRF are adopted as the training set, while Part_A is used as the test set.

Three representative methods, i.e., MCNN [69], CSRNet [20], and SCAR [4] are used as the competitors. Comparative results are reported in Table 6. It proves that the FPANet outperforms all competitors in terms of MAE and RMSE, which verifies the generalization ability of the proposed method.

5.8 Cross-domain analysis

To further validate the generalization ability of the proposed method, we perform cross-domain analysis on two car crowd datasets, namely CARPK and PUCPR+ [9]. The CARPK dataset includes 89,777 cars in various scenes for 4 different parking lots, while the PUCPR+ dataset consists of 17,000 cars in total. We followed the standard evaluation protocol in the corresponding benchmark to evaluate the performance [9].

Table 7 exhibits the scores of MAE and RMSE between the proposed FPANet and some state-of-the-art vehicle counting methods [9, 22, 30, 38, 46, 48, 62]. Comparative results demonstrate that the proposed model is flexible for various detection and counting tasks, and it consistently performs better than other competitors. Some Visualization

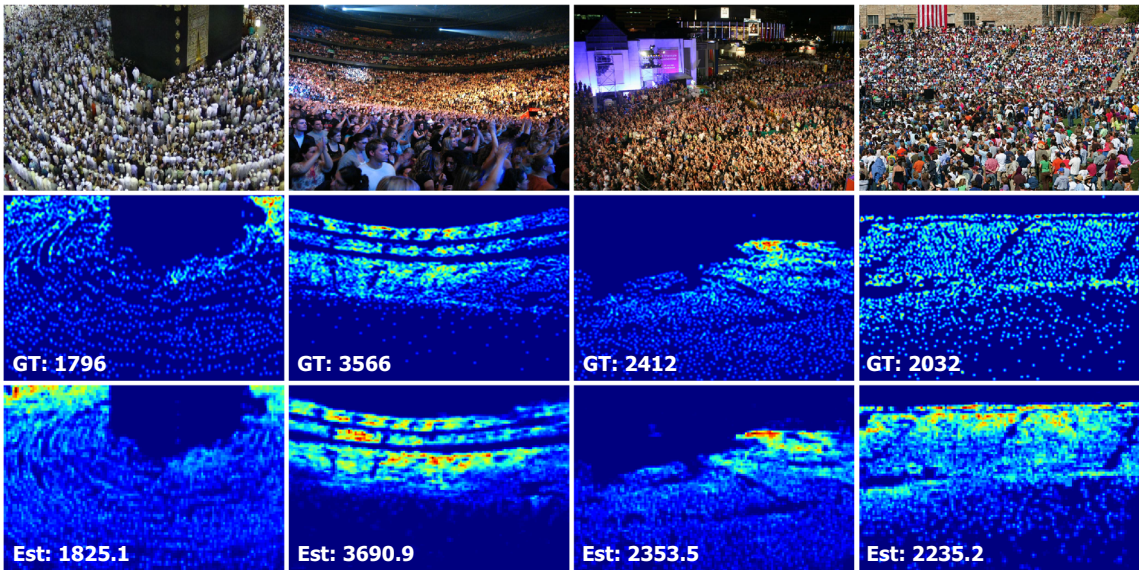


Fig. 9 Visualization results on UCF-QNRF dataset. The first row represents the four samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. The results prove that the proposed FPANet performs well in extremely dense scenarios with nonuniform background illumination

results on the CARPK and PUCPR+ datasets are illustrated in Fig. 11.

5.9 Efficiency comparison

To verify the efficiency of the proposed FPANet, several comparative methods are carried out to measure the parameters and complexity of the calculations. The input size is set to 576 × 768. The comparison results are illustrated in Table 8. Comparative results prove that the proposed method FPANet can achieve the best values with 59.9 and 7.8 in GFLOPs and parameters. Specifically, the FPANet reduces the GFLOPs and parameters by 71.9% and 61.8% compared with the DANet [3] which extracts the multiscale information to solve the scale variation. Compared with the GSANet [63] which employs the group convolution structure to reduce the computational cost, the proposed FPANet improves the GFLOPs by 10.2% and parameters by 10.1%.

Table 5 Experimental results on the NWPU-Crowd dataset (The best results are marked in **bold**)

Methods	MAE	RMSE
MCNN [69]	232.5	714.6
SANet [2]	190.6	491.4
A-CCNN [17]	176.5	520.6
ADMG [49]	152.8	907.3
STANet [60]	122.6	468.3
CRSNet [20]	121.3	378.8
PCC-Net [5]	112.3	457.0
SUA [34]	111.7	443.2
TopoCount [1]	107.8	438.5
BL [32]	105.4	454.2
SFCN [55]	105.7	424.1
KDMG [50]	100.5	415.5
DA^2Net [65]	102.6	378.5
FPANet(ours)	**97.1**	**372.8**

5.10 Ablation study

5.10.1 Ablation study on pivotal components

The validity of pivotal components in FPANet is verified on ShanghaiTech_Part A dataset by designing several step-wise models with various combinations. The detailed configurations are depicted as follows.

1. baseline: vanilla model without other components.
2. baseline+PSA: baseline with solely PSA (pyramid spatial attention) unit.
3. baseline+LCA: baseline with solely LCA (light-weight channel attention) unit.
4. baseline+PSA_LCA: baseline with the sequential connection of PSA and LCA units following the PSA-first and LCA-second order.
5. baseline+LCA_PSA: baseline with the sequential connection of PSA and LCA units following the LCA-first and PSA-second order.

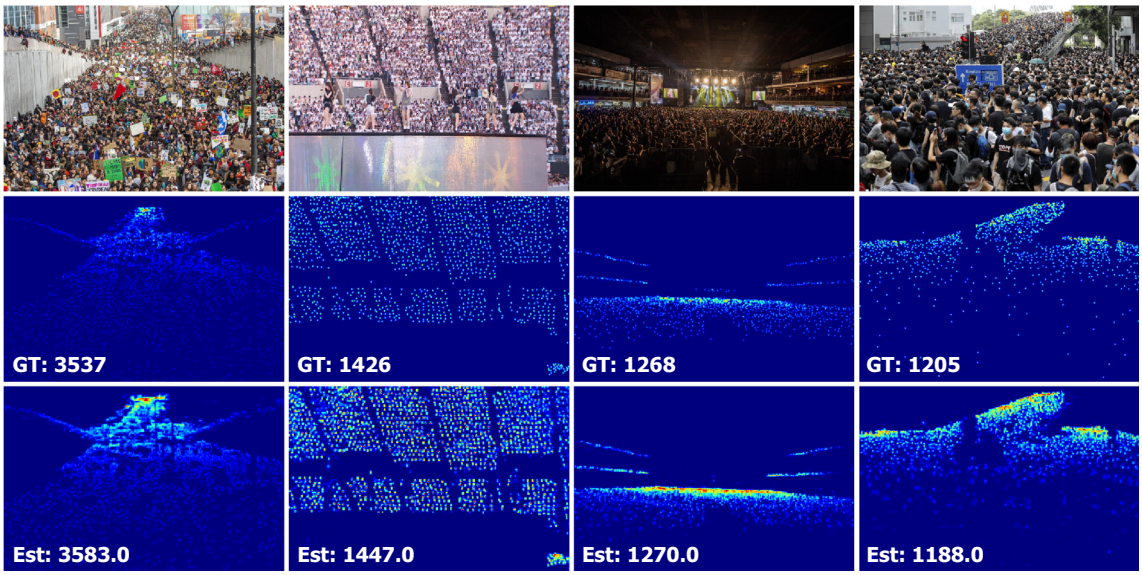


Fig. 10 Visualization results on NWPU-Crowd dataset. The first row represents the four samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. The visualization shows that the FPANet performs well on congested crowd scenes with a complicated background and large-scale variations

6. baseline+PSA‖LCA: baseline with parallel connection of the PSA and LCA units.
7. baseline+FP+PSA‖LCA+MA: baseline with FP (feature pyramid) module, attention module (parallel connection of the PSA and LCA units) and MA (multiscale aggregation) module.
8. baseline+FP+PSA_LCA+MA: final version, i.e., the proposed FPANet, the baseline with FP module, attention module (sequential connection of PSA and LCA units) and MA module.

The quantitative results of the step-wise models in terms of accuracy and efficiency are denoted in Table 9. It illustrates that all the pivotal components, i.e., feature pyramid (FP) module, attention module, i.e., pyramid spatial attention (PSA) unit and light-weight channel attention (LCA) unit, and multiscale aggregation (MA) module, facilitate the effective promotion in MAE and RMSE of the baseline method. The PSA unit outperforms the LCA unit in improving accuracy, but worse in improving efficiency. When the PSA and LCA units are sequentially connected (i.e., 'baseline+PSA_LCA' method and 'baseline+LCA_PSA' method) and concatenated in parallel (i.e., 'baseline+PSA‖LCA' method), the values of MAE and RMSE reduce evidently compared with the baseline combined single PSA or LCA units. Compared with 'baseline+LCA_PSA' method and 'baseline+PSA‖LCA' method, the 'baseline+PSA_LCA' performs better in improving the accuracy. Specifically, it reduces by 8.1% and 11.0% in MAE, and meanwhile reduces by 4.9% and 6.2% in RMSE, respectively. However, the accuracy is increased at the expense of efficiency. To address this problem, the FP and MA modules are equipped to reduce the values of GFLOPs and Params. Compared with the 'baseline+PSA_LCA', the 'baseline+FP+PSA_LCA+MA' reduces the value of GFLOPs by 8.4%, and Params by 54.7%, respectively. It outperforms all the other counterparts. Furthermore, the final method scores 70.9 in MAE and 120.6 in RMSE, both outperforming other ensemble methods in accuracy. This can be attributed to the multiscale

Table 6 Experimental results on the cross-data testing (The best results are marked in **bold**)

Methods	Part_A→Part_B		Part_A→QNRF		Part_B→Part_A		QNRF→Part_A	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [69]	73.7	101.6	340.3	375.2	182.1	264.5	233.7	404.9
CSRNet [20]	16.1	27.9	193.1	375.2	118.9	181.3	87.6	130.3
SCAR [4]	28.8	42.0	262.9	499.8	150.8	225.5	151.1	221.4
FPANet(Ours)	**11.7**	**20.9**	**171.9**	**329.9**	**105.8**	**170.9**	**76.0**	**120.0**

Table 7 Experimental results on the CARPK and PUCPR+ datasets (The best results are marked in **bold**)

Methods	CARPK		PUCPR+	
	MAE	RMSE	MAE	RMSE
FRCN [38]	74.4	82.3	109.2	144.5
IEP [46]	51.8	–	15.17	–
LPN [9]	23.8	36.8	22.8	34.5
SSD [30]	28.2	23.3	32.9	42.1
RetinaNet [22]	16.6	22.3	24.6	33.1
SCRDet [62]	11.1	25.4	9.1	13.5
FCOS [48]	10.7	13.6	16.0	23.8
FPANet(Ours)	**9.9**	**13.3**	**1.9**	**3.0**

Table 8 Comparison results of the FPANet and other methods in calculations and parameters (The best results are marked in **bold**)

Methods	GFLOPs	Params (M)
BL [32]	182.2	21.5
DSNet [3]	213.3	20.7
SFCN [55]	274.1	38.6
CSRNet [20]	182.7	16.3
SCAR [4]	182.9	16.3
GSANet [63]	66.7	8.7
FPANet (Ours)	**59.9**	**7.8**

feature extraction of FP module and the cross-hierarchy feature fusion MA module, which facilitates the exchange of feature map of different channel dimensions.

The qualitative comparisons of the baseline with different components are illustrated in Fig. 12. The exemplar images are affected by scale variances and background clustering, as illustrated in Fig. 12(a) and (b) is the ground truth. Figure 12(c) indicates that the estimated number and the density map of the baseline deviate the ground truth to a large extent. The PSA guarantees the accurate location of heads, as depicted in Fig. 12(d). The LCA can alleviate the error estimation for background regions, as depicted in Fig. 12(e). Both the compound modes of 'baseline+PSA_LCA' (Fig. 12(f)) and 'baseline+LCA_PSA' (Fig. 12(g)) boost the estimation accuracy, with the former being more effective. The 'baseline+PSA||LCA' (Fig. 12(h)) makes the problem even worse. Furthermore, the 'baseline+FP+MA' (Fig. 12(i)) performs effective effect. The final method (Fig. 12(j)) performs best in generating the density map and estimating the counting number.

5.10.2 Ablation study on scale size

The multiscale feature is extracted by the FP module and the PSA module. To explore the optimal scale size for the two modules, ablation studies are conducted on scale size, and the experimental results are presented in Tables 10 and 11.

As illustrated in Table 10, the FP module adopts the configuration with four scale sizes, i.e., $3 \times 3, 5 \times 5, 7 \times 7$ and 9×9. The final configuration, i.e., baseline+FP(3,5,7,9) outperforms the other configurations.

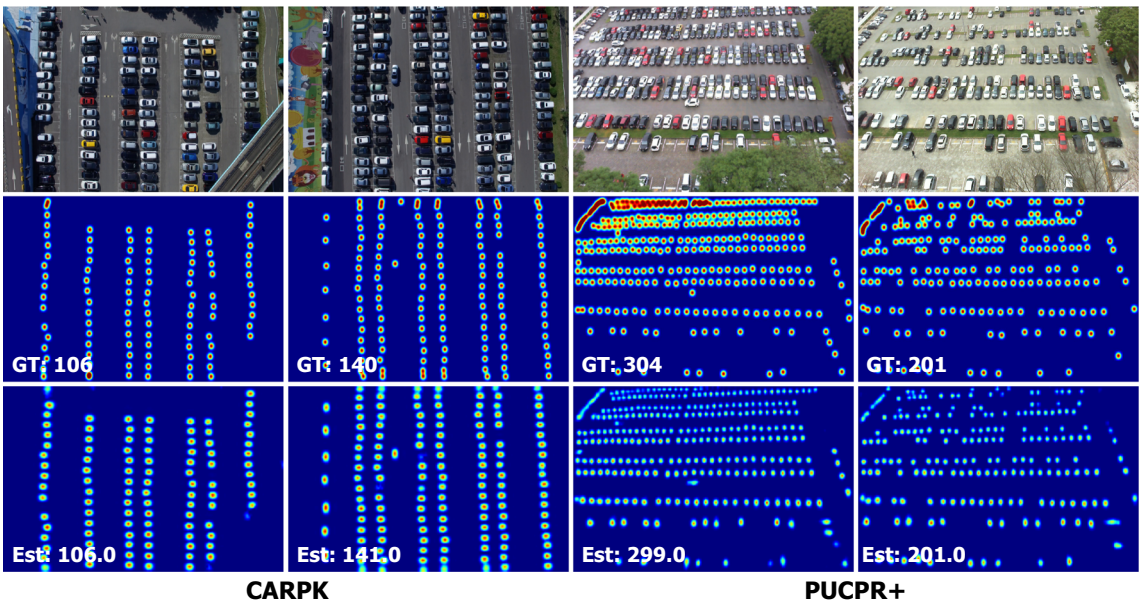
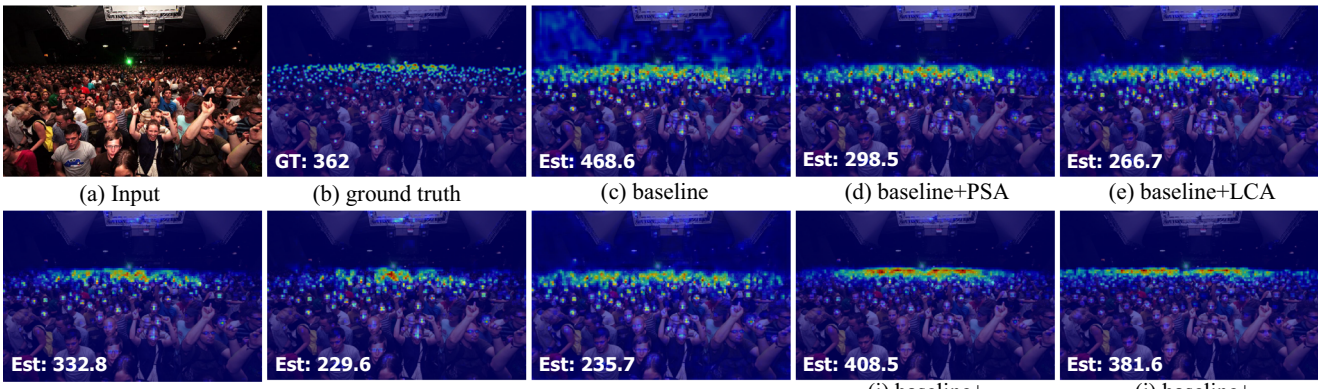


Fig. 11 Visualization results on CARPK dataset and PUCPR+ dataset. The first row represents the four samples. The second row shows the corresponding ground truth maps with real values. The last row illustrates the estimated maps with predicted values. It shows that the FPANet can attain predominant results in vehicle counting, with the estimated density map and counting number being very close to the ground truth

Table 9 Ablation analysis of the step-wise models in FPANet

Methods	Accuracy		Efficiency	
	MAE	RMSE	GFLOPs	Params (M)
Baseline	88.4	146.5	41.646	5.136
Baseline+PSA	79.0	134.9	66.243	17.325
Baseline+LCA	79.7	138.6	66.150	8.675
Baseline+PSA_LCA	76.7	123.4	66.286	17.525
Baseline+LCA_PSA	83.5	129.8	66.286	17.525
Baseline+PSA‖LCA	86.2	131.5	66.473	17.625
Baseline+FP+PSA‖LCA+MA	74.6	125.6	60.919	8.043
Baseline+FP+PSA_LCA+MA	**70.9**	**120.6**	**59.929**	**7.768**

The best results are marked in **bold**. For the indexes of GFLOPs and Params, the baseline scores the lowest value because it equips no component



(a) Input (b) ground truth (c) baseline (d) baseline+PSA (e) baseline+LCA

(f) baseline+ PSA_LCA (sequential) (g) baseline+ LCA_PSA (sequential) (h) baseline+ PSA‖LCA (parallel) (i) baseline+ FP+PSA‖LCA+MA (parallel) (j) baseline+ FP+PSA_LCA+MA (sequential)

Fig. 12 Qualitative comparisons on the step-wise models

Table 10 Comparisons of the FP module with different scale sizes (The best results are marked in **bold**)

Methods	MAE	RMSE
Baseline	88.4	146.5
Baseline+FP(3)	82.2	142.4
Baseline+FP(3,5)	78.9	133.1
Baseline+FP(3,5,7)	77.5	128.9
Baseline+FP(3,5,7,9)	**75.1**	**124.5**

Table 11 Comparisons of the PSA module with different scale sizes (The best results are marked in **bold**)

Methods	MAE	RMSE
Baseline	88.4	146.5
Baseline+PSA(1)	85.8	143.2
Baseline+PSA(1,2)	87.6	136.6
Baseline+PSA(1,2,4)	**79.0**	**134.9**

As depicted in Table 11, the PSA module employs three scale sizes, which utilize the adaptive average pooling operation with 4×4, 2×2, and 1×1. The final configuration, i.e., baseline+PSA(1,2,4), performs best compared with other configurations.

6 Conclusion

In this paper, we propose a Feature Pyramid Attention Network (FPANet) for accurate and efficient crowd counting. The FPANet consists of three modules, namely the feature pyramid module, attention module, and multiscale aggregation module. The feature pyramid module extracts multiscale features from the crowd to increase multiscale expression ability. The attention module consists of two cascaded attention units. It focuses on the head regions and handles the relations between channel maps, which restrains the background information. The multiscale aggregation module fuses diversified scale features and cross-channel attention information to enable information communication. Meanwhile, the FPANet adopts a multi-group structure to facilitate network efficiency. Comparative experiments on five benchmark crowd datasets and two cross-domain datasets have proven the superiority of the proposed FPANet compared with the state-of-the-art methods in terms of accuracy, efficiency, and generalizability.

Acknowledgements 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).

Declarations

The authors have no conflict of interest to declare that are relevant to the content of this article.

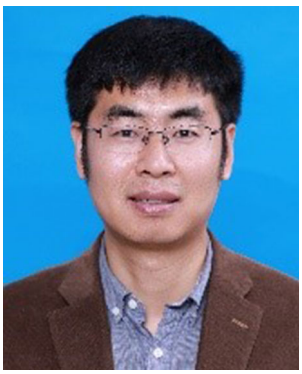
References

1. Abousamra S, Hoai M, Samaras D, Chen C (2021) Localization in the crowd with topological constraints. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 872–881. https://doi.org/10.1609/aaai.v35i2.16170
2. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750. https://doi.org/10.1007/978-3-030-01228-1_45
3. Dai F, Liu H, Ma Y, Zhang X, Zhao Q (2021) Dense scale network for crowd counting. In: Proceedings of the 2021 international conference on multimedia retrieval, pp 64–72. https://doi.org/10.1145/3460426.3463628
4. Gao J, Wang Q, Yuan Y (2019) Scar: spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8. https://doi.org/10.1016/j.neucom.2019.08.018
5. Gao J, Wang Q, Li X (2020) Pcc net: perspective crowd counting via spatial convolutional network. IEEE Trans Circuits Syst Video Technol 30:3486–3498. https://doi.org/10.1109/TCSVT.2019.2919139
6. Guo D, Li K, Zha Z, Wang M (2019) Dadnet: dilated-attention-deformable convnet for crowd counting. In: Proceedings of the ACM international conference on multimedia (ACM MM), pp 1823–1832. https://doi.org/10.1145/3343031.3350881
7. Guo J, Ma X, Sansom A, McGuire M, Kalaani A, Chen Q, Tang S, Yang Q, Fu S (2020) Spanet: spatial pyramid attention network for enhanced image recognition. In: Proceedings of the IEEE international conference on multimedia and expo (ICME), pp 1–6. https://doi.org/10.1109/ICME46284.2020.9102906
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/cvpr.2016.90
9. Hsieh MR, Lin YL, Hsu WH (2017) Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the international conference on computer vision (ICCV), pp 4165–4173. https://doi.org/10.1109/ICCV.2017.446
10. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7132–7141. https://doi.org/10.1109/TPAMI.2019.2913372
11. Huang S, Li X, Zhang Z, Wu F, Gao S, Ji R, Han J (2018) Body structure aware deep crowd counting. IEEE Trans Image Process 27:1049–1059. https://doi.org/10.1109/TIP.2017.2740160
12. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2547–2554. https://doi.org/10.1109/CVPR.2013.329
13. Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision (ECCV), pp 532–546. https://doi.org/10.1007/978-3-030-01216-8_33
14. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the international conference on international conference on machine learning (ICML), pp 448–456. https://doi.org/10.5555/3045118.3045167
15. Jiang G, Peng J, Wang H, Mi Z, Fu X (2022) Tensorial multi-view clustering via low-rank constrained high-order graph learning. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2022.3143848
16. Jiang X, Zhang L, Xu M, Zhang T, Lv P, Zhou B, Yang X, Pang Y (2020) Attention scaling for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4705–4714. https://doi.org/10.1109/cvpr42600.2020.00476
17. Kasmani SA, He X, Jia W, Wang D, Zeibots M (2018) A-ccnn: adaptive ccnn for density estimation and crowd counting. In: Proceedings of the IEEE international conference on image processing (ICIP), pp 948–952. https://doi.org/10.1109/ICIP.2018.8451399
18. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Proceedings of the international conference on learning representations (ICLR
19. Lei Y, Liu Y, Zhang P, Liu L (2021) Towards using count-level weak supervision for crowd counting. Pattern Recognit 109:107616. https://doi.org/10.1016/j.patcog.2020.107616
20. Li Y, Zhang X, Chen D (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), pp 1091–1100. https://doi.org/10.1109/CVPR.2018.00120
21. Li Z, Liu H, Zhang Z, Liu T, Xiong NN (2021) Learning knowledge graph embedding with heterogeneous relation attention networks. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3055147
22. Lin TY, Goyal P, Girshick RB, He K, Dollár P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42:318–327. https://doi.org/10.1109/TPAMI.2018.2858826
23. Liu H, Fang S, Zhang Z, Li D, Lin K, Wang J (2021a) Mfdnet: collaborative poses perception and matrix fisher distribution for head pose estimation. IEEE Trans Multimed 24:2449–2460. https://doi.org/10.1109/TMM.2021.3081873
24. Liu H, Zheng C, Li D, Shen X, Lin K, Wang J, Zhang Z, Zhang Z, Xiong NN (2021b) Edmf: efficient deep matrix factorization with review feature learning for industrial recommender system. IEEE Trans Industr Inf 18(7):4361–4371. https://doi.org/10.1109/TII.2021.3128240
25. Liu H, Liu T, Chen Y, Zhang Z, Li YF (2022a) Ehpe: skeleton cues-based gaussian coordinate encoding for efficient human pose estimation. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2022.3197364
26. Liu H, Zheng C, Li D, Zhang Z, Lin K, Shen X, Xiong NN, Wang J (2022b) Multi-perspective social recommendation method with graph representation learning. Neurocomputing 468:469–481. https://doi.org/10.1016/j.neucom.2021.10.050
27. Liu J, Gao C, Meng D, Hauptmann A (2018) Decidenet: counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5197–5206. https://doi.org/10.1109/CVPR.2018.00545
28. Liu L, Jiang J, Jia W, Amirgholipour S, Wang Y, Zeibots M, He X (2021c) Denet: a universal network for counting crowd with varying densities and scales. IEEE Trans Multimed 23:1060–1068. https://doi.org/10.1109/TMM.2020.2992979
29. Liu T, Wang J, Yang B, Wang X (2021d) Ngdnet: nonuniform gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom. Neurocomputing 436:210–220. https://doi.org/10.1016/j.neucom.2020.12.090
30. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu CY, Berg AC (2016) Ssd: single shot multibox detector. In: Proceedings of the European conference on computer vision (ECCV), pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
31. Ma YJ, Shuai HH, Cheng WH (2021) Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2021.3050059
32. Ma Z, Wei X, Hong X, Gong Y (2019) Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the international conference on computer vision (ICCV), pp 6141–6150. https://doi.org/10.1109/ICCV.2019.00624
33. Marsden M, McGuinness K, Little S, O'Connor N (2017) Fully convolutional crowd counting on highly congested scenes. In: Proceedings of the international joint conference on computer vision, imaging and computer graphics theory and applications (VISIGRAPP), pp 27–33. https://doi.org/10.5220/0006097300270033
34. Meng Y, Zhang H, Zhao Y, Yang X, Qian X, Huang X, Zheng Y (2021) Spatial uncertainty-aware semi-supervised crowd counting. In: Proceedings of the international conference on computer vision (ICCV), pp 15549–15559. https://doi.org/10.1109/ICCV48922.2021.01526
35. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the international conference on international conference on machine learning (ICML), pp 807–814. https://doi.org/10.5555/3104322.3104425
36. hwan Oh M, Olsen P, Ramamurthy K (2020) Crowd counting with decomposed uncertainty. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 11799–11806. https://doi.org/10.1609/AAAI.V34I07.6852
37. Ranjan V, Le HM, Hoai M (2018) Iterative crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp 278–293. https://doi.org/10.1007/978-3-030-01234-2_17
38. Ren S, He K, Girshick RB, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
39. Sajid U, Sajid H, Wang H, Wang G (2020) Zoomcount: a zooming mechanism for crowd counting in static images. IEEE Trans Circuits Syst Video Technol 30(10):3499–3512. https://doi.org/10.1109/TCSVT.2020.2978717
40. Sam DB, Babu RV (2018) Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 7323–7330. https://doi.org/10.1609/aaai.v32i1.12290
41. Sam DB, Surya S, Babu RV (2017) Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4031–4039. https://doi.org/10.1109/CVPR.2017.429
42. Sam DB, Peri SV, Sundararaman MN, Kamath A, Babu RV (2021) Locate, size, and count: accurately resolving people in dense crowds via detection. IEEE Trans Pattern Anal Mach Intell 43:2739–2751. https://doi.org/10.1109/tpami.2020.2974830
43. Shi Z, Zhang L, Liu Y, Cao X, Ye Y, Cheng MM, Zheng G (2018) Crowd counting with deep negative correlation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5382–5390. https://doi.org/10.1109/CVPR.2018.00564
44. Sindagi V, Patel V (2017a) Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6. https://doi.org/10.1109/AVSS.2017.8078491
45. Sindagi V, Patel V (2017b) Generating high-quality crowd density maps using contextual pyramid cnns. In: Proceedings of the international conference on computer vision (ICCV), pp 1879–1888. https://doi.org/10.1109/ICCV.2017.206
46. Stahl T, Pintea SL, Gemert JCV (2019) Divide and count: generic object counting by image divisions. IEEE Trans Image Process 28:1035–1044. https://doi.org/10.1109/TIP.2018.2875353
47. Thanasutives P, Ichi Fukui K, Numao M, Kijsirikul B (2021) Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In: Proceedings of the international conference on pattern recognition (ICPR), pp 2382–2389. https://doi.org/10.1109/ICPR48806.2021.9413286
48. Tian Z, Shen C, Chen H, He T (2019) Fcos: fully convolutional one-stage object detection. In: Proceedings of the international conference on computer vision (ICCV), pp 9626–9635. https://doi.org/10.1109/ICCV.2019.00972
49. Wan J, Chan AB (2019) Adaptive density map generation for crowd counting. In: Proceedings of the international conference on computer vision (ICCV), pp 1130–1139. https://doi.org/10.1109/ICCV.2019.00122
50. Wan J, Wang Q, Chan AB (2020) Kernel-based density map generation for dense object counting. IEEE Trans Pattern Anal Mach Intell:1–1. https://doi.org/10.1109/TPAMI.2020.3022878

51. Wang H, Peng J, Chen D, Jiang G, Zhao T, Fu X (2020a) Attribute-guided feature learning network for vehicle reidentification. IEEE MultiMed 27(4):112–121. https://doi.org/10.1109/MMUL.2020.2999464
52. Wang H, Peng J, Zhao Y, Fu X (2020b) Multi-path deep cnns for fine-grained car recognition. IEEE Trans Vehicular Technol 69(10):10484–10493. https://doi.org/10.1109/TVT.2020.3009162
53. Wang H, Wang Y, Zhang Z, Fu X, Zhuo L, Xu M, Wang M (2020c) Kernelized multiview subspace analysis by self-weighted learning. IEEE Trans Multimed 23:3828–3840. https://doi.org/10.1109/TMM.2020.3032023
54. Wang P, Gao C, Wang Y, Li H, Gao Y (2020d) Mobilecount: an efficient encoder-decoder framework for real-time crowd counting. Neurocomputing 407:292–299. https://doi.org/10.1016/j.neucom.2020.05.056
55. Wang Q, Gao J, Lin W, Yuan Y (2019a) Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 8190–8199. https://doi.org/10.1109/CVPR.2019.00839
56. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020e) Eca-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155
57. Wang Q, Gao J, Lin W, Li X (2021a) Nwpu-crowd: a large-scale benchmark for crowd counting and localization. IEEE Trans Pattern Anal Mach Intell 43:2141–2149. https://doi.org/10.1109/TPAMI.2020.3013269
58. Wang Q, Han T, Gao J, Yuan Y (2021b) Neuron linear transformation: modeling the domain shift for crowd counting. IEEE Trans Neural Netw Learn Syst:1–13. https://doi.org/10.1109/TNNLS.2021.3051371
59. Wang Y, Hu S, Wang G, Chen C, Pan Z (2019b) Multi-scale dilated convolution of convolutional neural network for crowd counting. Multimed Tools Appl 79:1057–1073. https://doi.org/10.1007/s11042-019-08208-6
60. Wen L, Du D, Zhu P, Hu Q, Wang Q, Bo L, Lyu S (2021) Detection, tracking, and counting meets drones in crowds: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7808–7817. https://doi.org/10.1109/CVPR46437.2021.00772
61. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
62. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, Sun X, Fu K (2019) Scrdet: towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the international conference on computer vision (ICCV), pp 8231–8240. https://doi.org/10.1109/ICCV.2019.00832
63. Zhai W, Gao M, Anisetti M, Li Q, Jeon S, Pan J (2022a) Group-split attention network for crowd counting. J Electr Imaging 31(4):041214. https://doi.org/10.1117/1.JEI.31.4.041214
64. Zhai W, Gao M, Souri A, Li Q, Guo X, Shang J, Zou G (2022b) An attentive hierarchy convnet for crowd counting in smart city. Cluster Comput:1–13. https://doi.org/10.1007/s10586-022-03749-2
65. Zhai W, Li Q, Zhou Y, Li X, Pan J, Zou G, Gao M (2022c) Da2net: a dual attention-aware network for robust crowd counting. Multimed Syst. https://doi.org/10.1007/s00530-021-00877-4
66. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 833–841. https://doi.org/10.1109/CVPR.2015.7298684
67. Zhang L, Shi M, Chen Q (2018) Crowd counting via scale-adaptive convolutional neural network. In: Proceedings of the IEEE workshop on applications of computer vision (WACV), pp 1113–1121. https://doi.org/10.1109/WACV.2018.00127
68. Zhang L, Shi Z, Cheng MM, Liu Y, Bian JW, Zhou JT, Zheng G, Zeng Z (2021) Nonlinear regression via deep negative correlation learning. IEEE Trans Pattern Anal Mach Intell 43:982–998. https://doi.org/10.1109/TPAMI.2019.2943860
69. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 589–597. https://doi.org/10.1109/CVPR.2016.70
70. Zou Z, Cheng Y, Qu X, Ji S, Guo X, Zhou P (2019) Attend to count: crowd counting with adaptive capacity multi-scale cnns. Neurocomputing 367:75–83. https://doi.org/10.1016/J.NEUCOM.2019.08.009

Wenzhe Zhai is pursuing the M.S. degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include smart city system, information fusion, crowd analysis and deep learning.



Mingliang Gao received his Ph.D. in Communication and Information Systems from Sichuan University. He is now an associate professor at the Shandong University of Technology. He was a visiting lecturer at the University of British Columbia during 2018-2019. He has been the principal investigator for a variety of research funding, including the National Natural Science Foundation, the China Postdoctoral Foundation, National Key Research Development Project, etc. His research interests include computer vision, machine learning, and intelligent optimal control. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley. He serves as a reviewer for more than 30 journals, e.g., Information Fusion, IEEE Transaction on Image processing, Pattern recognition, and IEEE Transactions on Instrumentation & Measurement.

Qilei Li is a third year Ph.D. student in Computer Science, Queen Mary University of London, supervised by Prof. Shaogang (Sean) Gong. Previously, he received the M.S. degree from Sichuan University in 2020. His research interests include computer vision and deep learning, particularly focusing on person ReID, video/image enhancement. He is a student member of IEEE, and he serves as a reviewer for Information Fusion, IEEE TIM, IEEE Access, Concurrency and Computation: Practice and Experience, and Multimedia System.

Gwanggil Jeon received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2003, 2005, and 2008, respectively. From 2009.09 to 2011.08, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. From 2011.09 to 2012.02, he was with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, as an Assistant Professor. From 2014.12 to 2015.02 and 2015.06 to 2015.07, he was a Visiting Scholar at Centre de Mathématiques et Leurs Applications (CMLA), École Normale Supérieure Paris-Saclay (ENS-Cachan), France. From 2019 to 2020, he was a Prestigious Visiting Professor at Dipartimento di Informatica, Università degli Studi di Milano Statale, Italy. He is currently a Full Professor at Incheon National University, Incheon, Korea. He was a Visiting Professor at Sichuan University, China, Universitat Pompeu Fabra, Barcelona, Spain, Xinjiang University, China, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, and University of Burgundy, Dijon, France. Dr. Jeon is an IEEE Senior Member, an Associate Editor of Sustainable Cities and Society, IEEE Access, Real-Time Image Processing, Journal of System Architecture, and MDPI Remote Sensing. Dr. Jeon was a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020.

Marco Anisetti is a full Professor at the Università degli Studi di Milano, Italy. He received the Ph.D degree in Computer Science from the Università degli Studi di Milano in 2009. He is the winner of the GIRPR award for the best Ph.D. thesis in 2010 and the winner of Chester Sall Award from IEEE Consumer Electronics Society in 2009. His research interests are in the area of Computational Intelligence and its application to the design and evaluation of complex systems and services. He is currently applying Big Data analytics to compute security and assurance metrics of Cloud systems and IoT systems in order to verify compliance to standards and policies. He has published several papers in journals and conference proceedings, and has served in the program committee of several international conferences.