Cluster Computing (2023) 26:1099–1111
https://doi.org/10.1007/s10586-022-03749-2

An attentive hierarchy ConvNet for crowd counting in smart city

Wenzhe Zhai[1] · Mingliang Gao[1] · Alireza Souri[1,2] · Qilei Li[3] · Xiangyu Guo[1] · Jianrun Shang[1] · Guofeng Zou[1]

Received: 14 April 2022 / Revised: 19 July 2022 / Accepted: 6 September 2022 / Published online: 22 September 2022

Abstract
Crowd counting plays a crucial rule in the development of smart city. However, the problems of scale variations and background interferences degrade the performance of the crowd counting in real-world scenarios. To address these problems, a novel attentive hierarchy ConvNet (AHNet) is proposed in this paper. The AHNet extracts hierarchy features by a designed discriminative feature extractor and mines the semantic features in a coarse-to-fine manner by a hierarchical fusion strategy. Meanwhile, a re-calibrated attention (RA) module is built in various levels to suppress the influence of background interferences, and a feature enhancement (FE) module is built to recognize head regions at various scales. Experimental results on five people crowd datasets and two cross-domain vehicle crowd datasets illustrate that the proposed AHNet achieves competitive performance in accuracy and generalization.

Keywords Smart city · Crowd counting · Attention mechanism · Hierarchical strategy

1 Introduction

Crowd counting is a new and developing problem in computer vision. It has a wide range of applications in smart city, e.g., public safety, urban planning smart environment sensing, and smart homes [1, 38]. However, the problems of scale variation caused by perspective of camera, and the background interferences in complex scenes pose a great challenge to accurately predict head regions in dense crowd scenarios.

To address this problem, a plethora of approaches have been put forward in the literature [6, 27]. Early approaches mainly consist of detection-based methods and regression-based methods. Recently, convolutional neural networks (CNN) have been extensively explored and exhibited outstanding performance in crowded counting.

Specifically, some CNN-based methods employ multi-column with different larger kernel sizes to extract the scale-aware context, which are difficult to train and could lead to information redundancy [23, 24, 36]. To address these issues, several studies utilize a single column network to extract the fixed receptive field, but the diverse spatial context information can not be obtained [4, 7, 15, 25]. Moreover, some literatures adopt attention mechanism to improve the prediction accuracy [17, 37]. But they ignored the effect of scale information on the prediction of outcomes. Overall, the aforesaid approaches suffer from substantial accuracy degradation when applied to crowd scenarios with large-scale fluctuations and heavy congestion.

In this paper, the attentive hierarchy ConvNet (AHNet) is proposed to alleviate the negative effects of the scale variations and background interferences, and improve the performance of crowd counting in dense crowd scenarios. The AHNet consists of a discriminative feature extractor and a hierarchical feature aggregator. The discriminative feature extractor extracts multi-scale features under different receptive field. The hierarchical feature aggregator fuses multi-layer features with attention and scale information to generate high-quality density maps.

Specifically, the hierarchical feature aggregator contains a re-calibrated attention (RA) module and a feature

✉ Mingliang Gao
mlgao@sdut.edu.cn

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

[2] Department of Computer Engineering, Haliç University, 34394 Istanbul, Turkey

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

enhancement (FE) module. The RA module adopts a RA module to integrate the intra- and inter-channel information by aggregating the input features along the vertical and horizontal directions into two separate direction-aware feature maps. Thus, it can capture long-range dependencies of the input feature map to pay more attention to the head regions and suppress the background interferences. The FE module is built to extract more scale features by expanding the fields-of-view of each convolutional layer. Thus, it enables the network to recognize head regions at various scales. In sum, the contributions of the proposed AHNet are as follows.

(1) An RA module is built in various levels to enhance the features from hierarchical layers. Meanwhile, it incorporates the inter-channel relationships and captures intra-channel dependencies simultaneously, to suppress the influence of background interferences.
(2) An FE module is built to extract more scale features by expanding the fields-of-view of each convolutional layer. It enables the network to recognize head regions at various scales.
(3) With the help of RA and FE modules, the AHNet is proposed to address the problems of scale variations and background interferences, and promote the accuracy of counting in complex crowd scenarios.

The rest of the paper is described as follows. Section 2 provides an overview of the related works. The framework of AHNet is introduced in Sect. 3. The experimental results and discussions are explained in Sect. 4. The conclusion is described in Sect. 5.

2 Related work

Crowd counting has drawn much attention and yielded numerous effective methods in the past decades. Many algorithms have been proposed to deal with the crowd counting task, and they can be broadly categorized into traditional-based methods and CNN-based methods. The early traditional-based methods are based on detection and regression, which work well in crowd scenarios with low density. Recently, benefiting from the powerful learning ability of CNNs, Fu et al. [5] firstly put forward the CNN-based model to improve accuracy for crowd counting in dense crowd scenarios. Subsequently, the CNN-based methods have become the mainstream in the field of crowd counting. The readers are suggested to refer to some survey papers [6, 20, 26, 27] for more details about the problems and the state-of-the-art methods in this domain.

2.1 Traditional-based methods

Traditional counting approaches mainly concentrated on detection-based schemes and the regression-based counting methods. The detection-based approaches [30] utilized a sliding window-like detector to detect pedestrians and count the quantity in the crowd. They used low-level characteristics such as Haar wavelets, histograms of oriented gradients, and boundaries localized from the whole human body. However, their performance would degrade because of the occlusion and background clutter.

The regression-based counting methods [3, 11] improve the counting accuracy by learning a mapping between image features and the total count density map. For instance, Chan et al. [2] used Gaussian process regression to learn the relationship between the extracted features and the population counts. Chen et al. [3] introduced the concept of cumulative attributes to construct a regression model from sparse and unbalanced training data. Idrees et al. [11] estimated counts using a variety of sources, including texture repetition, low-confidence head identification, and frequency-domain analysis. Unfortunately, in highly dense scenarios, even the best-performing detection approaches and regression-based approaches are unable to meet the requirements of crowd counting tasks due to the problems of scale variations and background interferences.

2.2 CNN-based methods

CNN-based methods [20] have obtained remarkable progress in the crowd counting domain. Some methods extracted the different scale features using multi-column architectures with different kernel sizes. Zhang et al. [36] utilized a three-column CNN, named MCNN, which dealt with the scale variation in dense scenarios. Sam et al. [23] introduced switching multi-column architecture to extract the features among different scale. Shi et al. [24] extracted scale information utilizing three parallel filters. However, these methods are difficult to train, and could lead to a lot of information redundancy [6, 15].

To address these problems, many works adopt single deeper networks to acquire larger receptive fields. For example. Sindagi and Patel [25] adopted an end-to-end cascaded CNN which learned high-level global priors to help estimate density maps from images of large variations. Kasmani et al. [13] introduced a method which can generate the different scale density maps adaptively for various patches to estimate the corresponding density map. Gao et al. [7] proposed a multi-task perspective crowd counting network that encoded hierarchical features and perspective variations for the crowd scenarios. Sam et al. [22] presented an incrementally growing CNN that was replicated

into two child regressors to predict the suitable expert for a given test patch. Ding et al. [4] introduced an encoder–decoder CNN that used the feature maps from both the encoding and decoding subnetworks to produce a more faithful density map which can estimate the number of persons more reliable. Li et al. [15] adopted dilated convolution operation to increase the receptive field for boosting performance on crowd counting. Although the aforementioned deeper CNNs approaches are effective to solve the problems of scale variations and background interferences, they ignore the crowd attention information.

In addition, several approaches equipped with attention mechanism have been proposed and achieved great success. Liu et al. [17] utilized an attention mechanism to decide which detection and regression modules have more reliable results to select appropriate counting models at different locations. Zhu et al. [37] presented an attentive multi-stage CNN which fuses coarse-to-fine hierarchical information and soft attention information to make the network pay more attention to foreground information. Zou et al. [39] introduced the adaptive capacity multi-scale CNNs, which can assign different attention weights for high-density and sparse regions. Gao et al. [8] presented a spatial-/channel wise attention regression network to acquire the discriminative features of each channel and pixel-wise context information of each image for crowd counting.

3 Framework of AHNet

3.1 Network structure

The overall architecture of the AHNet is depicted in Fig. 1. It mainly contains a discriminative feature extractor and a hierarchical feature aggregator. The original crowd image is fed into VGG-16 as discriminative feature extractor, and it utilizes feature maps from multiple convolutional layers to encode multi-scale features. There are five feature levels with downsampling ratios of $\{2, 4, 8, 16, 32\}$ respectively. The corresponding feature maps are denoted as $\{I_1, I_2, I_3, I_4, I_5\}$. The feature map P_i in each decoding stage is generated by the combination of I_i and P_{i+1}.

The process of hierarchical feature aggregator is as follows. Firstly, the feature map P_i from a higher level is up-sampled using the nearest-neighbor interpolation to make it the same size as I_i. Then, the upsampled map is combined with I_i by channel-wise concatenation. Furthermore, a 3×3 convolutions with ReLU activation is adopted to build the final feature map P_i. The feature representations are denoted to $\{P_1, P_2, P_3, P_4\}$, which are generated hierarchically using the features of different levels in the fusion feature extractor. However, due to the limited receptive field of specific I_i, it is only suitable to predict heads within a narrow range of scales.

To take full advantage of multi-level features, an RA module and an FE module are built in hierarchical feature aggregator. The RA module enables the channel-wise enhancement with the goal of boosting the foreground at various levels. In addition, the FE module pays attention to relevant spatial locations in the feature maps. The final high-quality density map is produced by fusing features from different enhanced layers.

3.2 Re-calibrated attention module

The RA module can be regarded as a computational unit to improve the expressive capability of learned features. The architecture of RA module is depicted in Fig. 2. It takes the intermediate feature map $F = [f_1, f_2, \ldots, f_C] \in \mathbb{R}^{C \times H \times W}$ as input and generates an augmented feature $M = [m_1, m_2, \ldots, m_C]$ as output.

In order to encode the attention information, the majority of current CNN models produce a channel attention map using the global pooling [10]. However, the global pooling operation squeezes global spatial information into a channel description. Thus, it is hard to reserve the location information of the head. To address this problem, we take the feature map F as input and re-calibrate it along with the horizontal direction and vertical direction using two spatially pooling kernels $(H, 1)$ and $(1, W)$, respectively.

The outputs of the cth channel at height h and width w are denoted as,

$$\begin{aligned} g_c^h(h) &= \frac{1}{W} \sum_{0 \le i < W} f_c(h, i), \\ g_c^w(w) &= \frac{1}{H} \sum_{0 \le j < H} f_c(j, w). \end{aligned} \tag{1}$$

These two recalibration processes fuse the features in two spatial directions, and generate a pair of direction-aware feature maps. Thus, more attentions are paid on two spatial directions and the background information can be restrained. The two branches are concatenated as follows,

$$s = \delta(S([g^h, g^w])), \quad s \in \mathbb{R}^{C/r \times (H+W)}, \tag{2}$$

where δ denotes a non-linear activation function. $S(\cdot)$ represents the concatenation operation along the spatial dimension. s is a middle feature map encoding horizontal and vertical spatial information.

The feature map along the spatial dimension is split into two independent maps, i.e., $s^h \in \mathbb{R}^{C/r \times H}$ and $s^w \in \mathbb{R}^{C/r \times W}$, where r denotes the reduction ratio to control the block size. Subsequently, two 1×1 convolutional operations are utilized to retain the final output with the same input

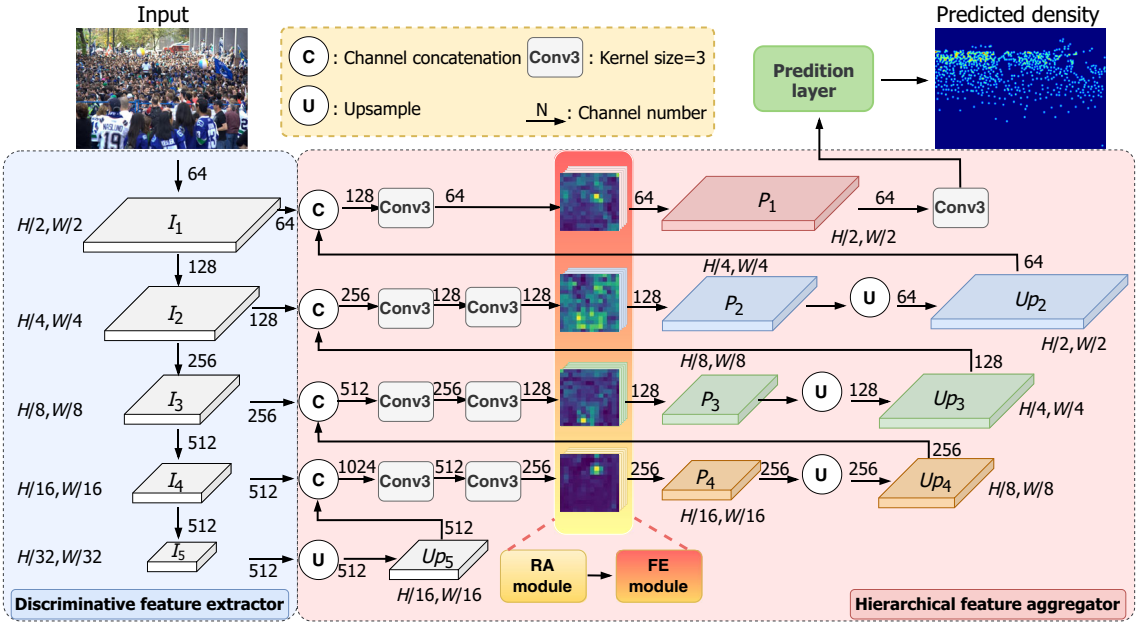

Fig. 1 Flowchart of the AHNet for crowd counting

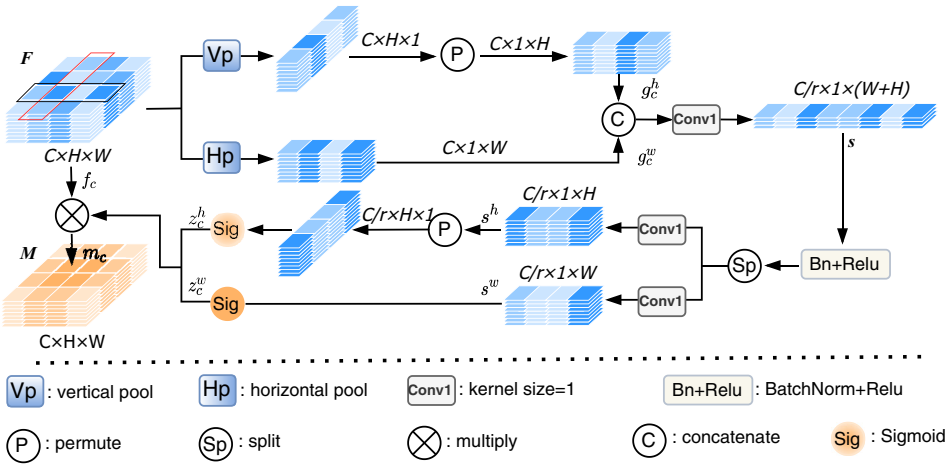

Fig. 2 The architecture of the RA module

channel number. The spatial attention weights are denoted as,

$$z^h = \sigma(\mathrm{Conv}_1(s^h)),$$
$$z^w = \sigma(\mathrm{Conv}_1(s^w)), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activate function. $\mathrm{Conv}_1(\cdot)$ indicates a convolution operation with the kernel size of 1×1. Finally, the output of the attention map M is presented as,

$$M = f_c(i,j) \times z^h(i) \times z^w(j). \tag{4}$$

3.3 Feature enhancement module

To further extract the scale information of the attention map, the FE module is built, as shown in Fig. 3. The attention map $M \in \mathbb{R}^{C\times H\times W}$ is divided into two feature maps, i.e., $M_1 \in \mathbb{R}^{C/2\times H\times W}$ and $M_2 \in \mathbb{R}^{C/2\times H\times W}$, which are transferred into two pathways for collecting different types of contextual information.

In the first pathway, M_1 generates the intermediate feature layer V_1 as,

$$V_1 = \mathrm{Up}\{\mathrm{Conv}_3[\mathrm{Avg}(M_1)]\}, \tag{5}$$

where $\mathrm{Up}(\cdot)$ represents the upsampling operation. $\mathrm{Conv}_3(\cdot)$ indicates a convolution operation with the kernel size of 3×3. $\mathrm{Avg}(\cdot)$ is the average pooling operation and it can

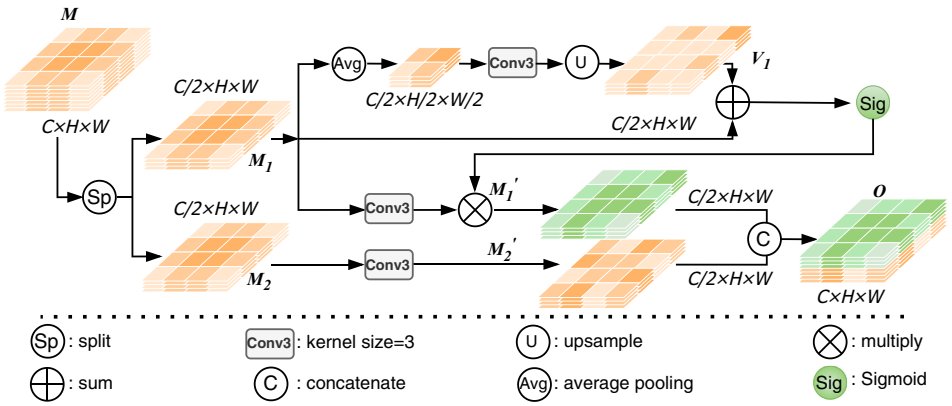

Fig. 3 The architecture of the FE module

change the size of the feature map. Avg$(\cdot)$ operation cooperates with Conv$_3(\cdot)$ operation to obtain the scale information from the attention map.

Then, the enhanced middle feature map of the first path can be formulated as,

$$M_1' = \sigma(M_1 \oplus V_1) \otimes \text{Conv}_3(M_1), \tag{6}$$

where σ denotes the sigmoid function. $\oplus$ and $\otimes$ denote the element-wise summation and multiplication, respectively.

In the second pathway, a convolution operation is performed to retain the original spatial context. It is formulated as,

$$M_2' = \text{Conv}_3(M_1). \tag{7}$$

Finally, the two pathway are concatenated by,

$$O = \text{Cat}(M_1', M_2'), \quad O \in \mathbb{R}^{C/G \times H \times W}, \tag{8}$$

where Cat$(\cdot)$ represents the concatenation operation.

3.4 Density map generation

Supposing the head coordinate as x_i, the head with an impulse function is formulated as $\delta(x - x_i)$. The whole heads of the image can be donated as $\sum_{i=1}^{N} \delta(x - x_i)$, where N represents the quantity of heads in the images. As the heads are dispersed, the Gaussian kernel is adopted to blur the labelled heads as follows,

$$M(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x), \tag{9}$$

where $M(x)$ is the density map. $*$ represents the convolution operation, and G_σ represents the Gaussian kernel. The crowd count is obtained by integrating over the density map.

3.4.1 Loss function

The network is trained by minimizing the Euclidean distance between the predicted density map and the ground truth density map,

$$\text{loss} = \frac{1}{M} \sum_{i=1}^{M} \| F_\theta(I_i) - Y_i \|_2^2, \tag{10}$$

where M indicates the amount of training samples. I_i represents the ith input image. $F_\theta(I_i)$ denotes the estimated crowd count. Y_i is the ith ground truth. $F_\theta(I_i)$ and Y_i employ the estimated crowd count and the ith ground truth, respectively.

4 Experimental results and analysis

4.1 Implementation details

The training and test are performed on an NVIDIA RTX3090 GPU with 24G memory in a PyTorch framework. The Adam optimization is adopted with the batch size of 6. According to the previous works [8, 25], the learning rate is initialized as 10^{-5} and reduces $\times 0.995$ per epoch. All the images and the corresponding density maps are resized to 576×768. The epoch number is set to 500.

4.2 Evaluation metric and datasets

The accuracy of the crowd counting evaluation is usually evaluated via the mean absolute error (MAE) and root mean squared error (RMSE), which are formulated as follows,

$$\text{MAE} = \frac{1}{N} \sum |c_i - \hat{c_i}|, \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum |c_i - \hat{c_i}|^2}, \tag{12}$$

where N is the amount of testing images. c_i and $\hat{c_i}$ are ground truth and the predicted values of the ith image, respectively.

4.3 Experimental analysis

The proposed method is evaluated on five benchmark datasets, including ShanghaiTech [36], UCF_CC_50 [11], UCF-QNRF [12], WorldExpo'10 [33] and NWPU-Crowd [31]. The experimental results are summarized in Table 1.

The ShanghaiTech dataset [36] contains two parts, namely Part_A and Part_B. The former part has 482 images (270 images for training, 30 images for validation, and 182 images for test). The latter part includes 716 images (360 images for training, 40 images for validation, and 316 images for test). In ShanghaiTech Part_A, the proposed AHNet scores 67.5 in MAE and 106.0 in RMSE, both ranking the second place among the competitors. Even though the performance is slightly lower than DENet [18], the proposed results are still very competitive. In ShanghaiTech Part_B, the proposed method achieve the lowest MAE and RMSE among the competitors. The proposed AHNet scores 7.7 and 11.9 in terms of MAE and RMSE, which both perform best. Especially, compared with the SCAR [8] which also adopts the attention mechanism, the AHNet reduces the MAE and RMSE by 18.9% and 21.7%. The improvement is primarily due to the fact that the proposed network not only contains attentional information extracted by the RA module but also obtains information at different scales through the FE module, which facilitates the network to generate more accurate prediction results.

The UCF_CC_50 dataset [11] incorporates 50 images are composed of dense crowd scenarios. Following the general principle [11], fivefold cross-validation strategy is adopted for performance evaluation. The proposed method scores 197.3 and 268.5 in MAE and RMSE, respectively. The AHNet ranks the first place compared with other state-of-the-art approaches. Compared with AMCNN [37], which adopted the hierarchical strategy to tackle the problem of scale variation in congested scenarios, the proposed model illustrates an improvement of 27.6% and 30.7% in MAE and RMSE. The reason is that the layers at each level in the AMCNN [37] network are independent of the other layers, while the AHNet fuses multi-layer feature with attention information and scale information.

The UCF-QNRF dataset [12] includes 1535 challenging images (1081 images for training, 120 images for validation, and 334 images for test). It varies within a large range in crowd counts (from 49 to 12,865). The proposed AHNet scores 108.2 and 186.8 in MAE and RMSE, both ranking the first place among the competitors. Compared with DENet [18], which utilizes dilated convolution and transposed convolution to solve the scale change problem, the proposed AHNet improves the score of MAE and RMSE by 10.6% and 8.9% as the AHNet not only utilizes FE module to focus on the relevant spatial locations at various

Table 1 Objective evaluation of the proposed method and the competitor in terms of MAE and RMSE

Method	Part_A		Part_B		UCF_CC_ 50		UCF-QNRF		WorldExpo10						NWPU-Crowd	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	S1	S2	S3	S4	S5	Avg	MAE	RMSE
MCNN [36]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	509.1	3.4	20.6	12.9	13.0	8.1	11.6	232.5	714.6
CMTL [25]	101.3	152.4	20.0	31.1	322.8	341.4	252.0	514.0	3.8	32.3	19.5	20.6	6.6	16.6	–	–
Switch-CNN [23]	90.4	135.0	21.1	30.1	318.1	439.2	228.0	445.0	4.4	15.7	10.0	11.0	5.9	9.4	–	–
DecideNet [17]	–	–	20.8	29.4	–	–	–	–	2.0	13.1	8.9	17.4	4.8	9.2	–	–
C-CNN [24]	88.1	141.7	14.9	22.1	–	–	–	–	3.8	20.5	8.8	8.8	7.7	9.9	–	–
A-CCNN [13]	85.4	124.6	19.2	31.5	367.3	423.7	–	–	–	–	–	–	–	–	176.5	520.6
SaCNN [34]	83.8	139.2	16.2	25.8	314.9	424.8	–	–	2.6	13.5	10.6	12.5	3.3	8.5	–	–
AMCNN [37]	76.1	110.7	15.3	27.4	272.5	387.5	–	–	–	–	–	–	–	–	–	–
PCCNet [7]	73.5	124.0	19.2	31.5	240.0	315.5	148.7	247.3	1.9	18.3	10.5	13.4	3.4	9.5	112.3	457.0
DNCL [35]	73.5	112.3	18.7	26.0	288.4	404.7	–	–	1.9	12.1	20.7	8.3	2.6	9.1	–	–
IG-CNN [22]	72.5	118.2	13.6	21.1	291.4	349.4	–	–	2.6	16.1	10.15	20.2	7.6	11.3	–	–
ACM-CNN [39]	72.2	103.5	17.5	22.7	291.6	320.9	–	–	2.4	**10.4**	11.4	15.6	3.0	8.56	–	–
FMLF [4]	69.8	114.7	10.2	14.9	271.3	376.3	–	–	2.8	12.1	9.4	15.6	3.5	8.68	–	–
CSRNet [15]	68.2	115.0	10.6	16.0	266.1	397.5	–	–	2.9	11.5	8.6	16.6	3.4	8.6	121.3	387.4
SCAR [8]	66.3	114.1	9.5	15.2	259.0	374.0	264.8	418.3	–	–	–	–	–	–	110.0	495.3
DENet [18]	**65.5**	**101.2**	9.6	15.4	241.9	345.4	121.0	205.0	2.8	10.7	8.6	15.2	3.5	8.2	–	–
AHNet (ours)	67.5	106.0	**7.7**	**11.9**	**197.3**	**268.5**	**108.2**	**186.8**	**1.4**	10.5	**8.4**	**8.0**	**2.5**	**6.16**	**100.2**	**364.1**

The best results are highlighted in bold

but also adopts the RA module to extract attention information to suppress the estimation error for background region.

The WorldExpo'10 dataset [33] contains 3380 frames in 103 scenes in the training set and 600 labelled frames from the remaining 5 scenes in the testing set. According to [33], the comparisons are performed on five scenes with a configured ROI. The compared value of MAE in these five scenes (i.e., s1, s2, s3, s4 and s5) and the average value (Avg) are shown in Table 1. It illustrates that the proposed AHNet outperforms other competitors in s1, s3, s4 and s5. For the second scene (s2), it ranks the second place and scores 10.5 which is slightly higher than ACM-CNN (10.4). However, in the index of average MAE, the AHNet scores 6.16 which surpasses the second-best method DENet by 24.9%.

The NWPU-Crowd dataset [31] is a newly released benchmark crowd dataset. It includes 5109 images with 2,133,375 head annotations, which are divided into three parts (3109 images for training, 500 images for validation, and 1500 images for test). As it has been released only recently, there are relative few comparative methods reported their results on this dataset. As shown in Table 1, the proposed AHNet scores 100.2 and 364.1 in terms of MAE and RMSE both performing best. Compared with PCCNet [7], which is also adopted the attention mechanism, the AHNet takes into account the hierarchy features and utilizes attention mechanism to further improve the estimated accuracy.

Some subjective results are depicted in Fig. 4. It shows that the estimated results of the proposed method are close to the ground truth in both sparse and congested crowd scenes.

4.4 Generalization ability analysis

4.4.1 Cross-dataset analysis

To validate the generalization ability of the proposed AHNet, we train the model on ShanghaiTech Part_A, and deploy the trained model on the Part_B and UCF-QNRF datasets. Comparative results with MCNN [36], CRSNet [15], and SCAR [8] are shown in Table 2. It depicts that the proposed AHNet model trained on ShanghaiTech Part_A scores 13.0 and 24.4 on ShanghaiTech Part_B dataset, and 165.2 and 334.3 on UCF-QNRF dataset in terms of MAE and RMSE. Compared with the three methods, the proposed method performs better performance on generalization.

4.4.2 Cross-domain analysis

To further verify the generalization ability of the proposed model, the cross-domain analysis is performed on two vehicle datasets, i.e., CARPK and PUCPR+. These two datasets composed of vehicle images from the drone view and high-rise buildings, respectively. The CARPK dataset contains 89,777 cars in various scenes from 4 different parking lots, while PUCPR+ dataset contains about 17,000 cars in total. To evaluate the counting performance, the evaluation protocol in their benchmark is adopted [14].

Table 3 reports the comparative results of the proposed method and other vehicle counting methods [9, 16, 19, 21, 28, 29, 32]. It shows that the proposed method scores 9.7 and 13.6 in MAE and RMSE on CARPK dataset, and 1.9 and 3.0 on PUCPR+ dataset, both outperforming the competitors. Some visualization results on CARPK and PUCPR+ are depicted in Fig. 5. It shows that the proposed method can achieve remarkable results in the domain of vehicle counting.

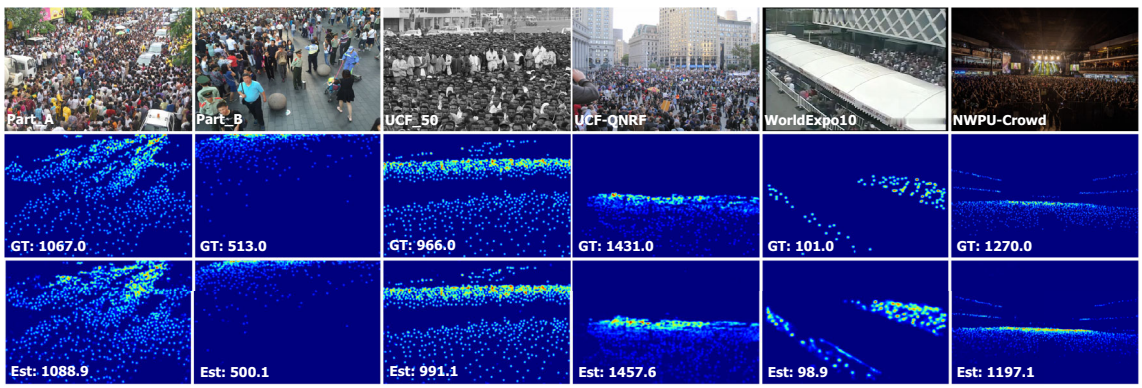

Fig. 4 Subjective evaluations on the benchmark datasets. The first row, the second row and the third row depict the exemplars, the ground truth and the estimated results, respectively

Table 2 Comparative results on the cross-data testing

Methods	Source dataset	Target dataset	MAE	RMSE
Part_B Cross-Dataset				
MCNN [36]	Part_A	Part_B	73.7	101.6
CSRNet [15]	Part_A	Part_B	16.1	27.9
SCAR [8]	Part_A	Part_B	28.8	42.0
AHNet (ours)	Part_A	Part_B	**13.0**	**24.4**
UCF-QNRF Cross-Dataset				
MCNN [36]	Part_A	UCF-QNRF	340.3	571.9
CSRNet [15]	Part_A	UCF-QNRF	193.1	375.2
SCAR [8]	Part_A	UCF-QNRF	262.9	499.8
AHNet (ours)	Part_A	UCF-QNRF	**165.2**	**334.3**

The best results are highlighted in bold

Table 3 Experimental results on the CARPK and PUCPR+ datasets

Methods	CARPK		PUCPR+	
	MAE	RMSE	MAE	RMSE
FRCN [21]	74.4	82.3	109.2	144.5
IEP [28]	51.8	–	15.17	–
LPN [9]	23.8	36.8	22.8	34.5
SSD [19]	28.2	23.3	32.9	42.1
RetinaNet [16]	16.6	22.3	24.6	33.1
SCRDet [32]	11.1	25.4	9.1	13.5
FCOS [29]	10.7	13.6	16.0	23.8
AHNet (ours)	**5.6**	**7.6**	**2.7**	**4.1**

The best results are highlighted in bold

4.5 Ablation study

To further demonstrate the effectiveness of each component proposed in the AHNet, two ablation studies are carried out. In the first study, the impact of the core components (i.e., RA module and FE module) are investigated on UCF_CC_50 dataset. The details are denoted as follows.

(i) Baseline: the discriminative feature extractor and the hierarchical feature aggregator of the first four layers.
(ii) Baseline + RA: the baseline model with single RA module.
(iii) Baseline + RA + FA: the baseline model with RA module and FE module.

The overall quantitative performance is shown in Table 4. It shows that both the 'baseline + RA' and the 'baseline + RA + FE' enhance the baseline, but the latter is better. The proposed RA module and FE module synergistically assist the AHNet better than a single RA module. In consequence, crowd counting can be divided into two processes. For the first process, the RA module takes advantages of both the horizontal re-calibrate and the vertical re-calibrate to amplify the input feature map data and suppresses the background disturbance. For the second process, the FE module is utilized to localize the crowd area and improving the accuracy. The final 'baseline + RA + FE' boosts the 'baseline' by 22.0% and 21.9% in terms of average MAE and RMSE, respectively.

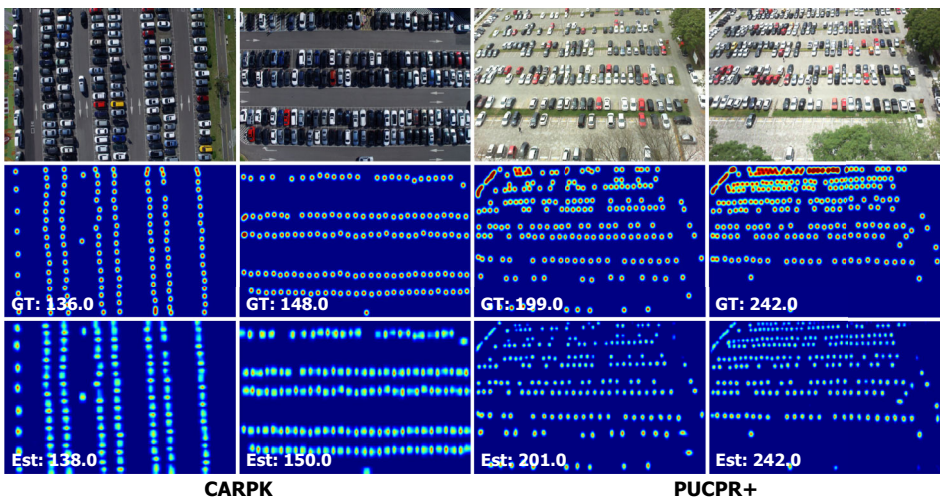

Fig. 5 An illustration of estimated density maps and crowd counts generated by the proposed method. The first row shows two set samples from CARPK datasets and PUCPR+ datasets. The second row shows the corresponding ground truth maps. The third row shows the density maps estimated by AHNet

Table 4 Effect of adopting the core components on crowd counting performance based on the UCF_CC_50 dataset

Methods	Metrics	Part 1	Part 2	Part 3	Part 4	Part 5	Avg
Baseline	MAE	205.5	256.5	334.2	184.5	285.1	253.1
	RMSE	252.8	338.9	467.9	435.2	225.6	344.1
Baseline + RA	MAE	175.7	232.3	297.4	180.9	278.6	232.9
	RMSE	232.7	305.0	371.3	227.0	465.5	320.3
Baseline + RA + FE	MAE	**139.5**	**176.8**	**273.7**	**173.3**	**223.2**	**197.3**
	RMSE	**193.0**	**212.4**	**358.3**	**210.9**	**367.9**	**268.5**

The best results are highlighted in bold

In the second study, the extensive ablation experiments are conducted to verify the effectiveness of the different level discriminative feature extractor and hierarchical feature aggregator. The details are denoted as follows.

(i) VGG-16: it indicates the original VGG-16 as a discriminative feature extractor without hierarchical feature aggregator.
(ii) VGG-16 + P_1: it includes discriminative feature extractor and the hierarchical feature aggregator of the first layer with RA module and FE module.
(iii) VGG-16 + P_2: it includes discriminative feature extractor and the hierarchical feature aggregator of the first two layers with RA module and FE module.
(iv) VGG-16 + P_3: it includes discriminative feature extractor and the hierarchical feature aggregator of the first third layers with RA module and FE module.
(v) VGG-16 + P_4 (w/o): it includes discriminative feature extractor and the hierarchical feature aggregator of the first fourth layers without RA module and FE module.
(vi) VGG-16 + P_4 (AHNet): the completed proposed method.
(vii) VGG-16 + P_5: it includes discriminative feature extractor and the hierarchical feature aggregator of the first fifth layers with RA module and FE module.

Table 5 Ablation analysis of the network with different architectures

Methods		MAE	RMSE
(i)	VGG16	77.5	131.7
(ii)	VGG16 + P_1	76.9	122.9
(iii)	VGG16 + P_2	75.2	124.4
(iv)	VGG16 + P_3	72.9	115.8
(v)	VGG16 + P_4 (w/o)	67.6	116.3
(vi)	VGG16 + P_4 (AHNet)	**67.5**	**106.0**
(vii)	VGG16 + P_5	67.7	112.6

The best results are highlighted in bold

Experimental results of the network with different architectures are shown in Table 5. Compared with the results of (i) and (ii), one can see that adding the first layer of hierarchical feature aggregator to the basic VGG-16 backbone can reduce the count error significantly. Compared with the results of (ii), (iii) and (iv), it proves that with the helps of RA and FE, the performance improves steadily along with the number of fusion layers. Meanwhile, by analysing the results of (v) and (vi), one can see that the proposed AHNet (with the RA module and FE module) produces better results than the architectures without these two modules. This demonstrates the effectiveness of the RA module and FE module. When the hierarchical feature aggregator combing the five layers [i.e., (vii)], the performance degrades to some extent. Thus, the architecture of "VGG16 + P_4" is chosen as the final AHNet.

The qualitative comparisons of the network with different architectures are shown in Fig. 6. The input 'exemplar' image suffers from background interferences and scale variations. From Fig. 6(i)–(iv), one can see that the visualized density map gradually changes from coarse-grained to fine-grained as the number of aggregator layers increases. In addition, as shown in the green box of Fig. 6(v) and (vi), the RA module and FE module can effectively suppress background interferences. By analysing the red box in Fig. 6(v) and (vi), one can see that the Fig. 6(v) is more close to the ground truth than Fig. 6(vi) which indicates that the combination of VGG16 and P4 (i.e., the proposed AHNet) outperforms the combination of VGG16 and P5.

4.6 Failure cases

Although the proposed AHNet demonstrates the superior performance in dense crowd scenarios, there are some unsatisfactory results in challenging scenes, as depicted in Fig. 7. When crowd scenes are acquired in low light, the estimated density maps are generated with a portion of unnecessary background interferences. Crowd counting in low-light environments is a technical challenge because the head region features are very close to the background region features in dim light environments. In future work,

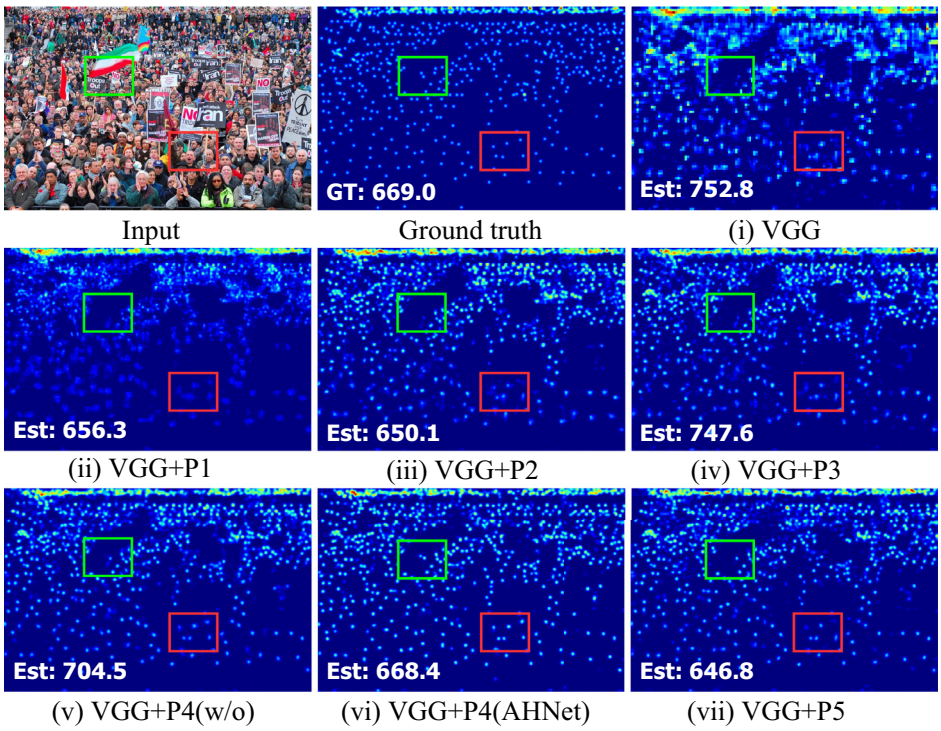

Fig. 6 Visualization of estimated results by the network with different architectures (the green and red boxes are two selected areas with background interferences and scale variations, respectively) (Color figure online)

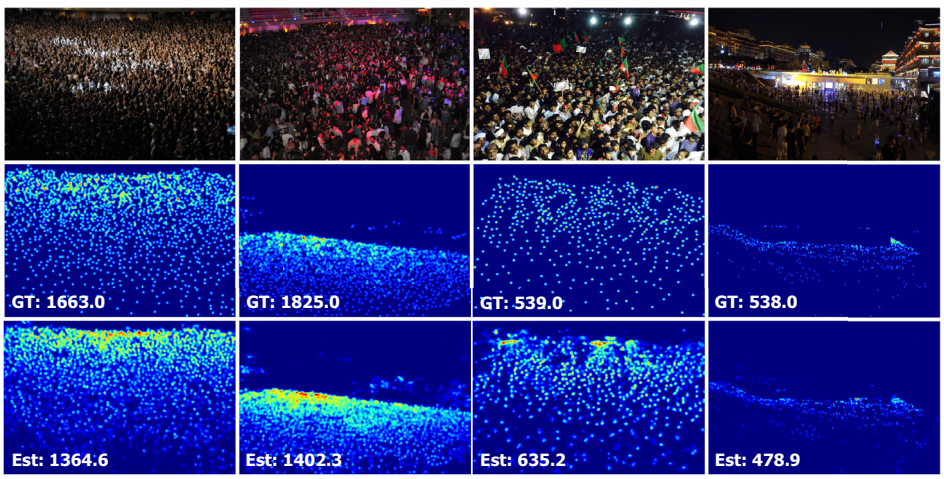

Fig. 7 The failure cases. The first row, the second row and the third row depict the exemplars, the ground truth and the estimated results, respectively

we further investigate reliable crowd feature map extraction in low-light environments.

5 Conclusion and future work

In this paper, an AHNet is proposed to address the problems of scale variations and background interferences in crowd counting and improve the performance of crowd counting in dense crowd scenarios. The hierarchical multi-scale features are extracted and integrated via the combination of the well-designed discriminative feature extractor and the hierarchical feature aggregator. Meanwhile, an RA module and an FE module are built and embedded in the hierarchical feature aggregator to suppress the background and recognize head regions at various scales. The proposed network is evaluated on five benchmark crowd datasets and two cross-domain vehicle crowd datasets. Experimental results demonstrate that the integration of the RA and FE modules boosts the baseline by 12.9% and 19.5% in terms of MAE and RMSE, respectively. Thus, the proposed AHNet performs favourably against the existing state-of-the-art methods. In future work, more research is expected to crowd counting in low-light scenes to improve the generalization ability in night environment.

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

References

1. Abualigah, L., Forestiero, A., Elaziz, M.A.: Bio-inspired agents for a distributed NLP-based clustering in smart environments. In: International Conference on Soft Computing and Pattern Recognition, 2021, pp. 678–687. Springer (2021). https://doi.org/10.1007/978-3-030-96302-6_64
2. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–7 (2008). https://doi.org/10.1109/CVPR.2008.4587569
3. Chen, K., Gong, S., Xiang, T., Loy, C.C.: Cumulative attribute space for age and crowd density estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2467–2474 (2013)
4. Ding, X., He, F., Lin, Z., Wang, Y., Guo, H., Huang, Y.: Crowd density estimation using fusion of multi-layer features. IEEE Trans. Intell. Transp. Syst. **22**, 4776–4787 (2021). https://doi.org/10.1109/TNNLS.2021.3084827
5. Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C.: Fast crowd density estimation with convolutional neural networks. Eng. Appl. Artif. Intell. **43**, 81–88 (2015). https://doi.org/10.1016/j.engappai.2015.04.006
6. Gao, G., Gao, J., Liu, Q., Wang, Q., Wang, Y.: CNN-based density estimation and crowd counting: a survey (2020). ArXiv: abs/2003.12783
7. Gao, J., Wang, Q., Li, X.: PCC Net: perspective crowd counting via spatial convolutional network. IEEE Trans. Circuits Syst. Video Technol. **30**, 3486–3498 (2020). https://doi.org/10.1109/TCSVT.2019.2919139
8. Gao, J., Wang, Q., Yuan, Y.: SCAR: spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing **363**, 1–8 (2019). https://doi.org/10.1016/j.neucom.2019.08.018
9. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the International Conference on Computer Vision (ICCV), 2017, pp. 4165–4173 (2017)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141 (2018). https://doi.org/10.1109/TPAMI.2019.2913372
11. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2547–2554 (2013). https://doi.org/10.1109/CVPR.2013.329
12. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 532–546 (2018). https://doi.org/10.1007/978-3-030-01216-8_33
13. Kasmani, S.A., He, X., Jia, W., Wang, D., Zeibots, M.: A-CCNN: adaptive CCNN for density estimation and crowd counting. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2018, pp. 948–952 (2018).https://doi.org/10.1109/ICIP.2018.8451399
14. Kiliç, E., Ozturk, S.: An accurate car counting in aerial images based on convolutional neural networks. J. Ambient Intell. Humaniz. Comput. (2021). https://doi.org/10.1007/s12652-021-03377-5
15. Li, Y., Zhang, X., Chen, D.: CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1091–1100 (2018). https://doi.org/10.1109/CVPR.2018.00120
16. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 318–327 (2020)
17. Liu, J., Gao, C., Meng, D., Hauptmann, A.: 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), 2018, pp. 5197–5206 (2018). https://doi.org/10.1109/CVPR.2018.00545
18. Liu, L., Jiang, J., Jia, W., Amirgholipour, S., Wang, Y., Zeibots, M., He, X.: DENet: a universal network for counting crowd with varying densities and scales. IEEE Trans. Multimed. **23**, 1060–1068 (2021). https://doi.org/10.1109/TMM.2020.2992979
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C.Y., Berg, A.C.: SSD: single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2
20. Nguyen, V., Ngo, T.D.: Single-image crowd counting: a comparative survey on deep learning-based approaches. Int. J. Multimed. Inf. Retr. **9**, 63–80 (2019). https://doi.org/10.1007/s13735-019-00181-y
21. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2015). https://doi.org/10.1109/TPAMI.2016.2577031
22. Sam, D.B., Sajjan, N.N., Babu, R.V.: Divide and grow: capturing huge diversity in crowd images with incrementally growing CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3618–3626 (2018)
23. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4031–4039 (2017). https://doi.org/10.1109/CVPR.2017.429
24. Shi, X., Li, X., Wu, C., Kong, S., Yang, J.S., He, L.: A real-time deep network for crowd counting. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2328–2332 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053780
25. Sindagi, V., Patel, V.: 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), 2017, pp. 1–6 (2017). https://doi.org/10.1109/AVSS.2017.8078491
26. Sindagi, V., Patel, V.: Generating high-quality crowd density maps using contextual pyramid CNNs. In: Proceedings of the International Conference on Computer Vision (ICCV), 2017, pp. 1879–1888 (2017). https://doi.org/10.1109/ICCV.2017.206
27. Sindagi, V.A., Patel, V.M.: A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognit. Lett. **107**, 3–16 (2018). https://doi.org/10.1016/j.patrec.2017.07.007
28. Stahl, T., Pintea, S.L., Gemert, J.C.V.: Divide and count: generic object counting by image divisions. IEEE Trans. Image Process. **28**, 1035–1044 (2019)

29. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the International Conference on Computer Vision (ICCV), 2019, pp. 9626–9635 (2019). https://doi.org/10.1109/ICCV.2019.00972
30. Viola, P.A., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**, 137–154 (2004)
31. Wang, Q., Gao, J., Lin, W., Li, X.: NWPU-Crowd: a large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 2141–2149 (2021). https://doi.org/10.1109/TPAMI.2020.3013269
32. Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K.: SCRDet: towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the International Conference on Computer Vision (ICCV), 2019, pp. 8231–8240 (2019). https://doi.org/10.1109/ICCV.2019.00832
33. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 833–841 (2015). https://doi.org/10.1109/CVPR.2015.7298684
34. Zhang, L., Shi, M., Chen, Q.: Crowd counting via scale-adaptive convolutional neural network. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV), 2018, pp. 1113–1121 (2018). https://doi.org/10.1109/WACV.2018.00127
35. Zhang, L., Shi, Z., Cheng, M.M., Liu, Y., Bian, J.W., Zhou, J.T., Zheng, G., Zeng, Z.: Nonlinear regression via deep negative correlation learning. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 982–998 (2021). https://doi.org/10.1109/TPAMI.2019.2943860
36. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589–597 (2016). https://doi.org/10.1109/CVPR.2016.70
37. Zhu, M., Wang, X., Tang, J., Wang, N., Qu, L.: Attentive multi-stage convolutional neural network for crowd counting. Pattern Recognit. Lett. **135**, 279–285 (2020)
38. Ziadeh, A., Abualigah, L., Elaziz, M.A., Şahin, C.B., Almazroi, A.A., Omari, M.: Augmented grasshopper optimization algorithm by differential evolution: a power scheduling application in smart homes. Multimed. Tools Appl. **80**(21), 31569–31597 (2021). https://doi.org/10.1007/s11042-021-11099-1
39. Zou, Z., Cheng, Y., Qu, X., Ji, S., Guo, X., Zhou, P.: Attend to count: crowd counting with adaptive capacity multi-scale CNNs. Neurocomputing **367**, 75–83 (2019). https://doi.org/10.1016/J.NEUCOM.2019.08.009

Wenzhe Zhai is pursuing his M.S. Degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include crowd counting and deep learning.

Mingliang Gao received his Ph.D. in Communication and Information Systems from Sichuan University, Chengdu, China, in 2013. He is currently an Associate Professor at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His main research interests include computer vision and deep learning.

Alireza Souri received his B.S. Degree in Software Engineering from the University College of Nabi Akram, Iran, and his M.Sc. and Ph.D. Degrees in Software Engineering from Science and Research Branch, Islamic Azad University, Iran. He is a Researcher and Lecturer at Islamic Azad University. Up to now, he has authored/co-authored 30 academic articles. He served on the program committees and the Technical Reviewer of several ISI-index journals and international conferences. He currently is an Associate Editor Member of Human-Centric Computing and Information Sciences (Springer), Cluster Computing (Springer) and IET Communications (IEEE) journals. His research interests include Formal Specification and Verification, Model checking, Grid and Cloud computing, IoT and Social networks. Now, He is a Member of The Society of Digital Information and Wireless Communications.

Qilei Li is currently a Ph.D. Student with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom. He received the M.S. Degree in Signal and Information Processing from Sichuan University. His research interests are computer vision and deep learning.

Xiangyu Guo is pursuing the M.S. Degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include computer vision and deep learning.

Jianrun Shang is pursuing the M.S. Degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include image enhancement and deep learning.

Guofeng Zou received Ph.D. in Pattern Recognition and Intelligent System from the College of Automation, Harbin Engineering University, Harbin, in 2013. He is currently working as a Lecturer in the College of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His current research interests include pattern recognition, digital image processing and analysis, and machine learning.