Group-split attention network for crowd counting

Wenzhe Zhai,[a] **Mingliang Gao**,[a,*] **Marco Anisetti**,[b] **Qilei Li**,[c] **Seunggil Jeon**,[d] **and Jinfeng Pan**[a]

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

[b]Universit'a degli Studi di Milano, Department of Computer Science, Milano, Italy

[c]Queen Mary University of London, School of Electronic Engineering and Computer Science, London, United Kingdom

[d]Samsung Electronics, Suwon, Republic of Korea

Abstract. Crowd counting is a considerable yet challenging task in intelligent video surveillance and urban security systems. The performance has been significantly boosted along with the springing up of the convolutional neural networks (CNNs). However, accurate and efficient crowd counting in congested scenes remains under-explored due to scale variation and cluttered background. To address these problems, we propose a biologically inspired crowd counting method named group-split attention network (GSANet). The GSANet consists of three principal modules, namely GS module, dual-aware attention module, and aggregation module. The GS module processes the subfeatures of each group in parallel, and groups the input feature map to reduce the computational cost. The dual-aware attention module synergies the spatial and channel dimensional information to alleviate the estimation error in background regions. The aggregation module adopts a learning-based cross-group strategy to aggregate and facilitate the fusion of feature maps along different channel dimensions. Extensive experimental results on five benchmark crowd datasets demonstrate that the GSANet achieves superior performances in terms of accuracy and efficiency. © *2022 SPIE and IS&T* [DOI: 10.1117/1.JEI.31.4.041214]

Keywords: crowd counting; convolutional neural network; attention mechanism; computer vision.

Paper 220257SS received Mar. 8, 2022; accepted for publication May 23, 2022; published online Jun. 6, 2022.

1 Introduction

The task of crowd counting is to estimate the number of people in crowded scenes. It has aroused much attention in recent years, as it plays a considerable role in many real-world applications, e.g., video surveillance, smart city governance, and public safety management.[1–3] However, scale variation and cluttered backgrounds in highly congested crowed scenes are still challenging in this domain.[4,5] The scale variations are caused by camera perspective distortion, which results in different distances between heads and cameras. And the cluttered background region (e.g., trees, buildings, vehicles, and so on) is similar to the foreground's region (e.g., head area) which leads to estimation errors in highly congested crowd scenes. The challenges in a congested scene are shown in Fig. 1.

To address the aforementioned problems, many efforts have been devoted[5] and many methods have been proposed. These methods can be divided into three categories, namely detection-based methods,[6–8] regression-based methods,[9,10] and deep learning-based method.[5,11] The detection-based methods[7,8] generally detect the person instances with pretrained detectors.[12,13] Although it performs well in sparse scenes, the estimation accuracy degrades to a large degree in congested scenes. The regression-based methods[9,10] build a mapping between low-level features and the crowd number. However, it tends to omit the location information, this is suboptimal for detection. Recently, benefitting from the powerful learning ability of convolutional neural networks (CNNs), the deep learning-based methods have play a dominant role in crowd

*Address all correspondence to Mingliang Gao, mlgao@sdut.edu.cn

(a) **Scale variation** (b) **Cluttered background**

Fig. 1 The challenges in congested scene. (a) Scale variation and (b) clusttered background.

counting.[4,14] The deep learning-based methods address this problem by estimating a density map, and estimate the number of crowds by integrating over the density map. As complementary, the attention mechanism has been adopted as a guidance to improve the counting accuracy.[15,16] Benefiting from the attention mechanism, the head region can gain more attention than the non-head region. Thus, the background disturbance can be suppressed. However, the deep learning-based models are extremely inefficient, which require significant computational cost and run at a low speed.[17,18] To solve this problem, some lightweight networks for crowd counting were proposed.[19,20] However, these methods struggle in seeking a satisfying trade-off between accuracy and efficiency.

In this paper, we put forward the group-split attention network (GSANet) for accurate and efficient crowd counting. The proposed method mainly consists of three modules, i.e., GS module, dual-aware attention module, and aggregation module. The GS module is designed to process the features of each group in parallel, which divides the input feature map into groups to reduce the computation cost. The attention module is designed in a dual-aware pattern, consisting of a spatial attention (SA) unit and a channel attention (CA) unit. The former unit concentrates on the head region, while the later unit guides the network to focus on the relation between channel maps to eliminate the error estimation for background. With the guidance of the attention module, the head region in the estimated density map gains more attention in encoding the final density map. Third, an aggregation module is built to fuse the multidimensional feature maps. Meanwhile, it cooperates with the GS module to reduce the computing time. To sum up, the contributions of our work are three aspects.

1. We design a GS module and an aggregation module to process the features of each group in parallel, which can reduce the computation cost and fuse multidimensional features, simultaneously.
2. We build an attention module in a dual-aware pattern, which gains more attention in the head region to copy with the problems of scale variation and cluttered background.
3. We carry out extensive experiments and ablation studies to verify the performance of counting in challenging scenarios and the effectiveness of the individual components in the proposed GSANet.

The rest of the paper is structured as follows. Section 2 presents an overview of the relevant work. Section 3 introduces the details of the proposed GSANet. Comparative results and discussions are presented in Sec. 4. This work is concluded in Sec. 5.

2 Related Work

2.1 *Detection-Based Method*

This method mainly employs a sliding-window-like detector scanning the image, and detecting the body or head of each person. Then, a classifier is trained to determine the positive instances.[21] Dollar[22] built a slide window detector over the image for crowd counting. Similarly, Li et al.[6] constructed detectors on the head and shoulder to estimate the number of people in the surveillance area. Recent approaches seek an end-to-end schema by CNN-based object detectors to improve the accuracy.[23,24] Although the detection-based method is successful in low-density scenes, the performance in highly congested environment is still unsatisfying.

2.2 *Regression-Based Method*

This approach aims to build a regression model to map the image characteristics to crowd number. These are feasible approaches in congested environments, as they don't require explicit pedestrian detection and segmentation. Davies et al.[25] pioneered the regression method in crowd counting by extracting the underlying features and building a linear regression model to map the features to the crowd number. Paszke et al.[26] adopted a random forest regression to learn the mapping of nonlinear features to people. Lempitsky and Zisserman[27] used spatial distribution information to regress a density map. However, these methods ignore the location information of each person, thus the practical application is limited.[28]

2.3 *Deep Learning-Based Method*

In recent years, benefitting from the strong ability of feature expression and computing resources, the deep learning has achieved great success in computer vision.[29–33] Zhang et al.[14] proposed a multicolumn CNN (MCNN) model to increase the receptive field to address the problem of scale variations. Similarly, a switch-CNN[4] was proposed through a switching mechanism multicolumn architecture to utilize the features among different scales. However, the multicolumn CNNs are usually difficult to train.[34,35] To address this problem, Li et al.[36] adopted a single column fully convolutional network[37] with cascaded dilated convolutional layers. Besides, many other architectures were designed to improve the estimated accuracy.

In addition, attention-based modules have been adopted in this domain to improve the estimation accuracy. Zhang et al.[38] proposed an attention model for crowd counting by estimating a probability map between the head areas and the nonhead areas. Hossain et al.[15] proposed a scale-aware network by combining both the global and local scale attentions.

Although the CNN-based methods achieve remarkable progress, their performance comes with the cost of burdensome computation. In this regard, how to reduce the number of parameters in the network draws much attention. For instance, Wang et al.[19] designed an encoder–decoder architecture with limited computation resources. Cao et al.[39] designed a scale aggregation network to improve the representation ability and scale diversity of density map in crowd counting. However, these approaches fail to find a balance between efficiency and accuracy. In this work, a GSANet is proposed to strike an optimal balance between accuracy and efficiency.

3 Proposed Method

3.1 *Overview*

As depicted in Fig. 2, the GSANet consists of three modules, i.e., GS module, dual-aware attention module, and aggregation module. First, following the general setting, a tailored ResNet-50 with the first three layers is used as the backbone network to extract features. Next, we adopt the group split module to process the features of each group in parallel and divide them into subgroups. Then, a dual-aware attention module consisting of an SA unit and CA unit is built upon

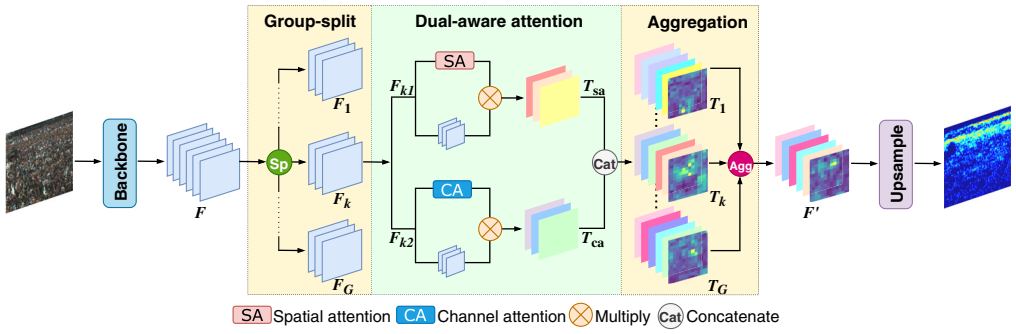

Fig. 2 Flowchart of the GSANet for crowd counting.

each group. Further, the two units are concatenated and transported to the aggregation module to fuse the subfeatures among different channel dimensions. Finally, an up-sample operation is applied to the output to predicate the density map.

3.2 *Group-Split Module*

The GS module takes the feature maps from the backbone as input and splits them into subgroups, as shown in Fig. 2. For a given feature map $F \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the channel number, height, and width, respectively. The GS module first divides F into G ($G = 32$ in this work) groups along the channel dimension

$$F = [F_1, F_2, \cdots, F_G], \quad F_k \in \mathbb{R}^{C/G\times H\times W}, \tag{1}$$

where each subfeature F_k captures a separated group unit response in the training process. The purpose of the splitting is to reduce the number of parameters. Then, the corresponding importance coefficient for each subfeature is generated by an attention module. Specifically, at the beginning of each attention unit, the input of F_k is split into two branches (e.g., F_{k1}, F_{k2}) along the channels dimension, i.e.

$$F_{k1}, F_{k2} = F_k/2, F_{k1}, \quad F_{k2} \in \mathbb{R}^{C/2G\times H\times W}. \tag{2}$$

3.3 *Dual-Aware Attention Module*

The dual-aware attention module composes an SA unit and a CA unit, which are organized in a parallel manner. The SA unit generates an SA map by utilizing the interspatial relationship of features. Meanwhile, the CA unit produces a CA map by exploiting the interaction among channels.

3.3.1 *Spatial attention unit*

The density map is generated by convolving the head location with a normalized Gaussian kernel to generate a smoother training gradient. (Note: more details are referred in Sec. 4.1.2.) The accurate location of heads is crucial to the generation of a density map. The spatial branch produces an SA map to emphasize or suppress features in different spatial locations.[40] To this aim, we build the SA unit to focus on the head region under the premise of ensuring accurate head detection. The architecture of the SA unit is shown in Fig. 3, and the spatial-enhanced feature map T_{sa} is formulated as follows:

$$T_{\text{sa}} = \sigma\{\text{C}_7[\text{MaxPool}(F_{k1}); \text{AvgPool}(F_{k1})]\} \otimes F_{k1}, \tag{3}$$

where the C_7 represents a convolution layer with the kernel size of 7×7. $\text{MaxPool}(\cdot)$ and $\text{AvgPool}(\cdot)$ denote the max pooling and average pooling on channel dimensions, respectively. $\sigma\{\cdot\}$ is the Sigmoid function. $\otimes$ denotes the element-wise multiplication.

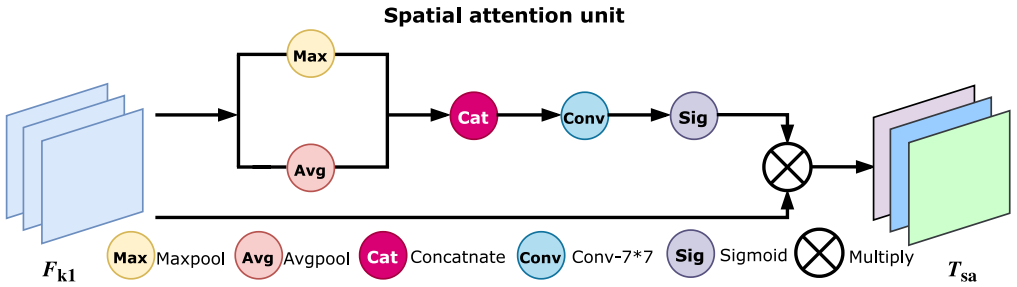

Fig. 3 The architecture of the SA unit.

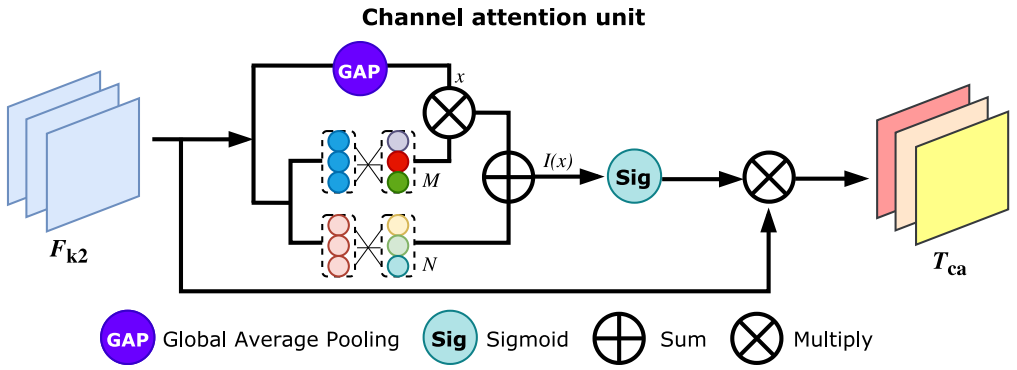

Fig. 4 The architecture of the CA unit.

3.3.2 *Channel attention unit*

The SA unit can result in error estimation for the background due to the resemblances between the foreground and background region texture. To address this problem, we design the complementary CA unit, as shown in Fig. 4. The CA unit is formulated as follows:

$$x = \mathcal{X}(F_{k2}) = \frac{1}{\mathrm{WH}} \sum_{i=1,j=1}^{W,H} F_{k2}(i,j), \quad x \in \mathbb{R}^{C/2G \times 1 \times 1}, \tag{4}$$

$$T_{ca} = \sigma\{\mathrm{I}(x)\} \otimes F_{k2} = \sigma(\mathrm{M}x + n) \otimes F_{k2}, \tag{5}$$

where $\mathcal{X}$ represents the channel-aware GAP which obtains the aggregated features of the background region. The CA unit generates channel weights (e.g., $M \in \mathbb{R}^{C/2G \times 1 \times 1}$ and $n \in \mathbb{R}^{C/2G \times 1 \times 1}$) by $\mathrm{I}(x)$. $\otimes$ denotes the element-wise multiplication. $T_{\mathrm{ca}} \in \mathbb{R}^{C/2G \times H \times W}$ is the enhanced feature map.

Then, the two branches are concatenated as follows:

$$T_i = \mathrm{Cat}(T_{\mathrm{sa}}, T_{\mathrm{ca}}), \quad T_i \in \mathbb{R}^{C/G \times H \times W}, \tag{6}$$

where $\mathrm{Cat}(\cdot)$ represents the concatenation operation, which fuses the spatial information and channel information of the C/G dimension.

3.4 *Aggregation Module*

The aggregation module adopts a learning-based cross-group strategy to aggregate and facilitate the exchange of feature map of different channel dimensions. It contains two operations, namely channel concatenation and shuffle.[41] The channel concatenation operation fuses the feature maps of different dimensions. It splices all the subfeatures and form the feature map $T \in \mathbb{R}^{C \times H \times W}$, which is formulated as

$$T = \Phi[T_1, T_2, \cdots, T_G], \quad T_k \in \mathbb{R}^{C/G \times H \times W}, \tag{7}$$

where $\Phi(\cdot)$ denotes the concatenate operation from the dimension of the G group channel. T_k is the k'th feature map.

The shuffle operation, as shown in Fig. 5, enables cross-group information communication along the channel dimension.[41] Specifically, it reassembles the feature map T to G groups whose output has C channels. The shuffle operation consists of three steps, as shown in Fig. 5. First, the input channel dimension is reshaped into feature tensor (G, C), and it is transposed to feature tensor (C, G). Then, the above output is flattened and divided into G group. Finally, the subgroups are spliced together to form the final new feature map F'.

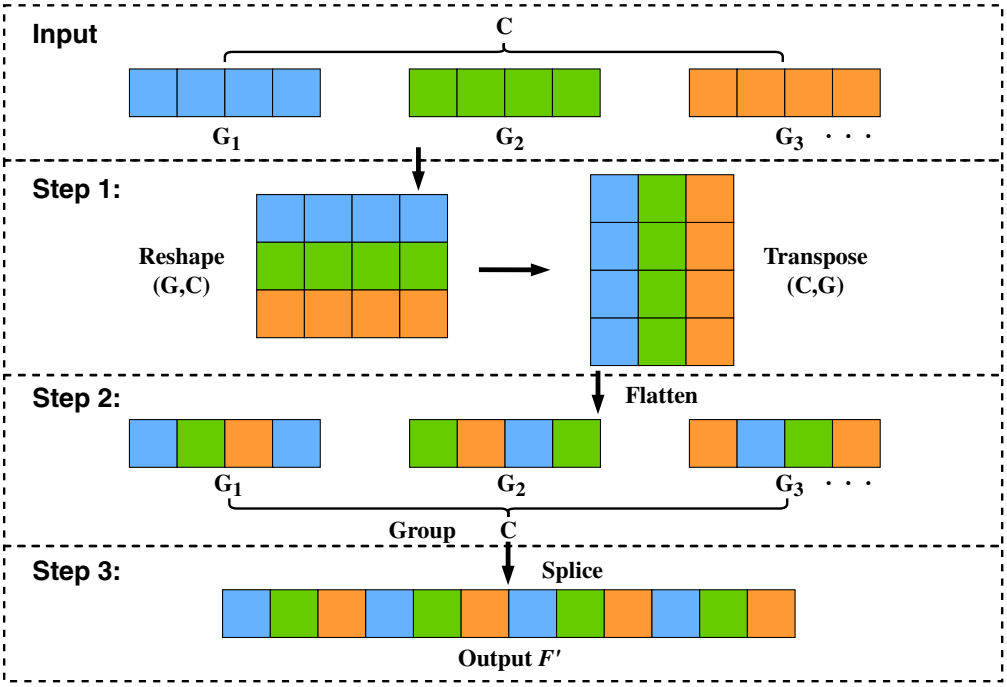

Fig. 5 Diagram of the shuffle operation.

4 Experimental Results and Analysis

4.1 *Implementation Details*

4.1.1 *Training details*

The training and test are performed on an NVIDIA RTX3090 GPU with the PyTorch framework.[42] All the images and the corresponding density maps are resized to 576×768. We adopt the Adam optimization. The learning rate is initialized as 10^{-5} and reduces $\times 0.995$ per epoch. We supplement the training detail with random clipping and horizontal flipping instead of vertical flipping, which minimizes overfitting and ensures the network is sufficiently trained.

4.1.2 *Density map generation*

To provide an efficient supervision for generating high-quality estimated crowd density maps, we transfer the labeled heads images to ground truth density maps. Following the work in Ref. 14, the density maps are generated as follows.

Suppose the head coordinate as x_i, we formulate the head with an impulse function 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 number of heads in the images. As the heads are dispersed, we adopt the Gaussian kernel to blur the labeled heads as follows:

$$M(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x), \tag{8}$$

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

4.1.3 *Evaluation metrics*

The mean absolute error (MAE) represents the sum of the absolute values of the differences between the predicted values and ground truth. It is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum |z_i - \hat{z}_i|, \tag{9}$$

where N represents the number of objects, z_i is the ground truth, and $\hat{z}_i$ is the predicted value.

The root-mean-square error (RMSE) represents the sum of the squares of the distances between predicted values and ground truth. It is formed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum |z_i - \hat{z}_i|^2}, \tag{10}$$

where the variables have the same meaning as Eq. (9).

4.1.4 *Loss function*

The proposed GSANet aims to optimize the loss function as follows:

$$\mathrm{loss} = \frac{1}{M}\sum_{i=1}^{M} \|F_\theta(I_i) - Y_i\|_2^2, \tag{11}$$

where M is the batch size. $F_\theta(I_i)$ indicates the estimated density map. I_i represents the input image. θ denotes the learned parameter, and Y_i is the density map of ground truth.

4.2 *Benchmark Datasets*

The comparison experiments are carried out on five challenging datasets (e.g., ShanghaiTech, UCF_CC_50, UCF-QRNF, WorldExpo'10, and NWPU-Crowd). The characteristics of the benchmark datasets are depicted in Table 1. The implementation details of crowd counting are illustrated and the performance of GSANet on these different datasets is compared with other competitors.

4.2.1 *Performance on ShanghaiTech dataset*

The ShanghaiTech dataset[14] consists of two parts, i.e., Part_A and Part_B, with a total number of 1198 images. There are a total of 1198 labeled images with 330,165 labeled heads. The Part_A dataset has 482 images, of which 300 are for training and the remaining 182 for test. The images in Part_A dataset are randomly crawled from the Internet, and they are across diverse scenes and largely varied densities. The Part_B dataset consists of 716 images, 400 images for training, and

Table 1 Characteristics of the benchmark datasets used for evaluation.

DataSet	Number of images	Min	Max	Average	Total
Part_A	482	33	3199	501	241,677
Part_B	716	9	578	124	88,488
UCF_CC_50	50	94	4543	1280	63,974
UCF-QNRF	1535	49	12,865	815	1,251,642
WorldExpo'10	3980	1	253	50	199,923
NWPU-Crowd	5109	0	20,033	418	2,133,375

Table 2 Experimental results on the ShanghaiTech dataset.

	Part_A		Part_B	
Method	MAE	RMSE	MAE	RMSE
Zhang et al.[43]	181.8	277.7	32.0	49.8
GP[44]	120.4	179.4	12.5	18.3
MCNN[14]	110.2	173.2	26.4	41.3
CMTL[45]	101.3	152.4	20.0	31.1
TDF-CNN[46]	97.5	145.1	20.7	32.8
NLT[47]	93.8	157.2	11.8	19.2
Switching-CNN[4]	90.4	135.0	21.1	30.1
DecideNet[16]	—	—	20.8	29.4
C-CNN[48]	88.1	141.7	14.9	22.1
AM-CNN[38]	87.3	132.7	15.6	26.4
SaCNN[49]	86.8	139.2	20.7	32.8
A-CCNN[50]	85.4	**124.6**	11.0	19.0
SAAN[15]	—	—	16.7	28.4
MATT[51]	80.1	129.4	11.7	17.5
GSANet (Ours)	**74.3**	130.3	**8.7**	**13.9**

Note: the best results are highlighted in **bold**.

316 for test. The images of Part_B are taken from the metropolis of Shanghai. By contrast, the images in Part_B have a smaller intradataset divergence. The subjective evaluation of the GSANet model with SOTA methods is reported in Table 2. In Part_A, the proposed method scores 74.3 and 130.3 in MAE and RMSE. Especially, it improves the MAE by 7.2% compared with the second-best method, MATT.[51] Meanwhile, it achieves a competitive performance in RMSE, ranking the third place among the competitors. In Part_B, the GSANet gains the results of 8.7 and 13.9 in MAE and RMSE, which are both the best results compared with other SOTA methods. The subjective results on the ShanghaiTech dataset are shown in Fig. 6. It proves that the estimated crowd density map can accurately depict the distribution of the crowd. Meanwhile, the estimated counting results are very close to the ground truth.

4.2.2 *Performance on UCF_CC_50 dataset*

The UCF_CC_50 dataset[10] includes 50 images in different resolutions. The crowd count per image varies from 94 to 4543. Intrinsically, not only the extremely-congested scenes but also the limited training samples cause this dataset being extremely challenging. The comparative results are presented in Table 3. The proposed method achieves the score of 166.4 in MAE and 235.8 in RMSE, both ranking the first place among all the SOTAs. Specifically, compared with ASNet[56] which also adopts the attention mechanism in crowd density estimation, the proposed GSANet reduces the score of MAE by 4.8%, and RMSE by 6.3%, respectively. Compared with PCC-Net[17] and MobileCount[19] which are both light-weight-based methods, the proposed GSANet reduces the MAE by 30.7% and 41.2%, and RMSE by 25.3% and 38.4%, respectively. The visualization of the estimated crowd density maps with counting number is depicted in Fig. 7. It proves that the estimated crowd density maps and counting number are approximate to the ground truth.

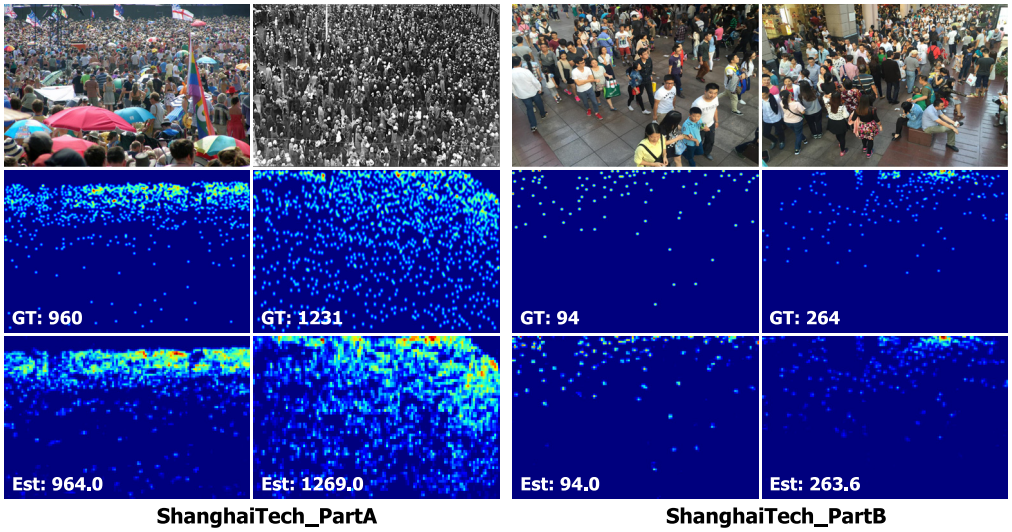

Fig. 6 The estimated density maps with counting number on exemplar images from the ShanghaiTech datasets. Note: the first, second, and third rows represent the input image, the ground truth, and the estimated results, repectively.

Table 3 Experimental results on the UCF_CC_50 dataset.

Methods	MAE	RMSE
Idrees et al.[10]	419.5	541.6
Zhang et al.[43]	467.0	498.5
MCNN[14]	377.6	509.1
MATT[51]	355.0	550.2
MobileCount[19]	283.1	382.6
CSRNet[36]	266.1	397.5
SCAR[52]	259.0	374.0
HA-CNN[34]	256.2	348.4
PCC-Net[17]	240.0	315.5
CAT-CNN[28]	235.5	324.8
LSC-CNN[53]	225.6	302.7
PFDNet[54]	205.8	289.3
D2CNet[55]	182.1	254.9
ASNet[56]	174.8	251.6
GSANet (Ours)	**166.4**	**235.8**

Note: the best results are highlighted in **bold**.

4.2.3 *Performance on UCF-QNRF dataset*

The UCF-QNRF[57] is a high-density and cross-scene dataset. It contains 1535 images with 1,251,642 annotations. Particularly, it has a wider variety of scenes, compared with other datasets. Following the criterion raised in Ref. 57, 1201 images are used for training and 334 images for test. Comparative results are shown in Table 4. It shows that the GSANet scores 110.0 in

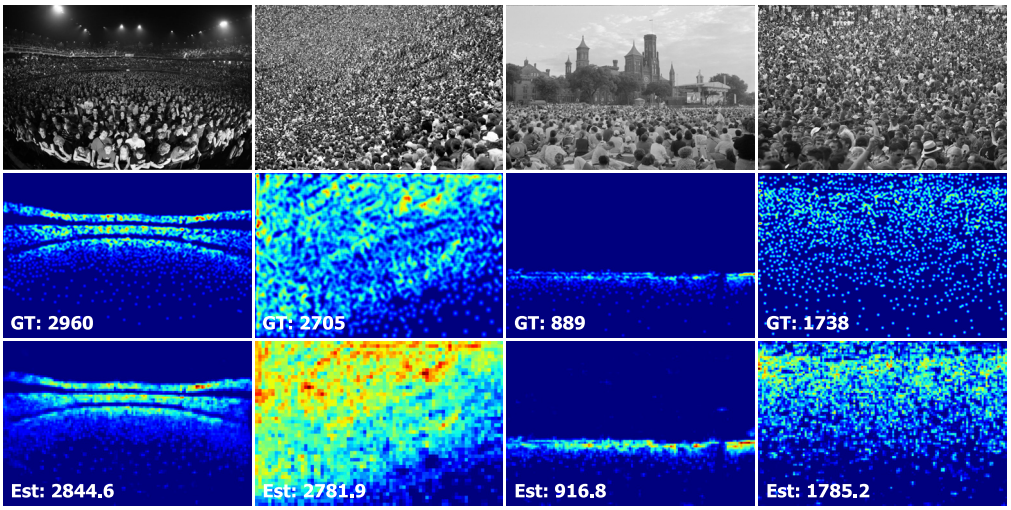

Fig. 7 The estimated density maps with counting number on exemplar images from the UCF_50 dataset. Note: the first, second, and third rows represent the input image, the ground truth, and the estimated results, respectively.

Table 4 Experimental results on the UCF-QNRF dataset.

Methods	MAE	RMSE
Zhang et al.[43]	467.0	498.5
Idress et al.[10]	315.0	508.0
MCNN[14]	277.0	509.1
SCAR[52]	264.8	418.3
Switching-CNN[4]	228.0	445.0
NLT[47]	172.3	263.1
PCCNet[17]	148.7	247.3
CRSNet[36]	129.0	209.0
DENet[58]	121.0	205.0
LSC-CNN[53]	120.5	218.2
DUBNet[59]	116.0	**178.0**
HA-CNN[34]	118.1	180.4
DADNet[60]	113.2	189.4
DA^{2}Net[61]	111.7	204.3
GSANet (Ours)	**110.0**	195.0

Note: the best results are highlighted in **bold**.

MAE which ranks the first place, and 195.0 in RMSE which ranks the fourth place. Particularly, it reduces the MAE and RMSE by 1.5% and 4.5% compared with DA^{2}Net,[61] which also leverages attention module. The experimental results indicate that the proposed method is superior to other methods in MAE, and remain competitive in RMSE. The estimated density maps with counting number on sample images from the UCF-QNRF dataset are shown in Fig. 8. It proves that both the estimated crowd density map and counting number are very closely related to the ground truth.

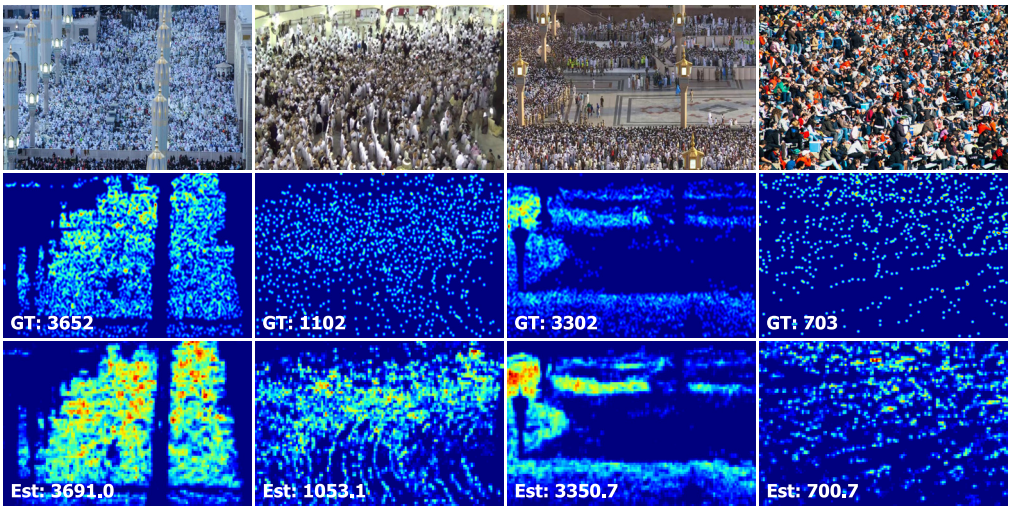

Fig. 8 The estimated density maps with counting number on exemplar images from the UCF-QNRF dataset. Note: the first, second, and third rows represent the input image, the ground truth, and the estimated results, respectively.

4.2.4 *Performance on WorldExpo'10 dataset*

The WorldExpo'10 dataset[43] is a cross-scene crowd counting dataset which consists of 3,980 frames with a total number of 199, 923 labeled pedestrians. Since five different regions of interest (ROI) and the perspective maps are provided for the test scenes (S1-S5), we count persons within the ROI area following the general criterion.[47,54] The performance of the proposed GSANet against the SOTA methods are shown in Table 5. It shows that the GSANet performs

Table 5 Experimental results on the WorldExpo'10 dataset.

Methods	S1	S2	S3	S4	S5	MAE (Avg.)
NLT[47]	2.3	22.8	16.7	19.7	3.9	13.1
Zhang et al.[43]	9.8	14.1	14.3	22.4	3.7	12.9
MCNN[14]	3.4	20.6	12.9	13.0	8.1	11.6
MSCNN[62]	7.8	15.4	14.9	11.8	5.8	11.7
SCAR[52]	1.9	13.8	9.6	29.8	3.9	11.8
DCL[63]	1.8	16.2	9.2	25.0	2.8	11.0
DecideNet[16]	2.0	13.1	8.9	17.4	4.8	9.2
CSRNet[36]	2.9	11.5	8.6	16.6	3.4	8.6
SANet[39]	2.6	13.2	9.0	13.3	3.0	8.2
DENet[58]	2.8	10.7	8.6	15.2	3.5	8.2
LSC-CNN[53]	2.9	11.3	9.4	12.3	4.3	8.0
STDNet[64]	1.8	12.8	10.3	7.8	**2.5**	7.04
EPF[65]	2.1	10.9	8.5	**5.4**	3.0	6.58
GSANet (Ours)	**1.5**	**10.5**	**8.0**	8.0	**2.5**	**6.1**

Note: the best results are highlighted in **bold**.

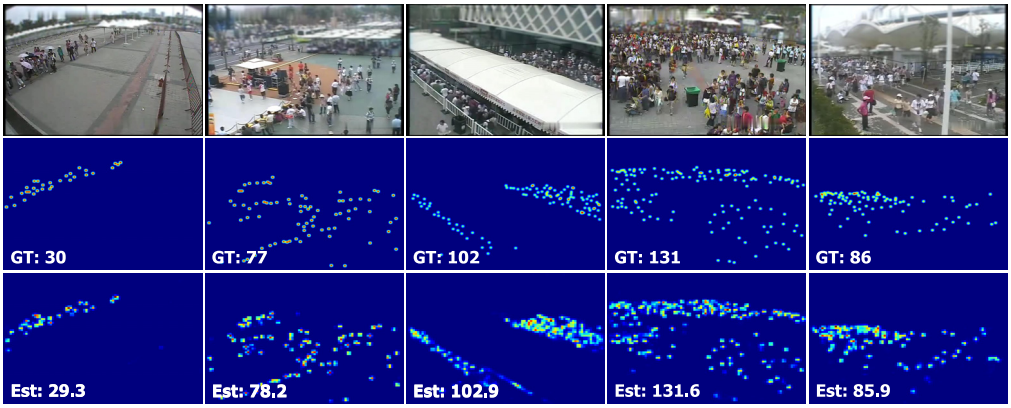

Fig. 9 The estimated density maps with counting number on exemplar images from the WorldExpo'10 dataset. Note: the first, second, and third rows represent the input image, the ground truth, and the estimated results, respectively.

best in Scenes 1, 2, 3, and 5, and ranks fourth in Scene 4. Specially, it ranks the first place in the average MAE and exceeds the second-best method EPF[65] by 7.3%. Figure 9 shows some estimated results in the WorldExpo'10 dataset. As shown in Fig. 9, the proposed method can accurately reflect the actual crowd distribution in all the images.

4.2.5 *Performance on NWPU-Crowd dataset*

The NWPU-Crowd dataset[66] is currently the largest congested crowd dataset. It contains 5109 images, in which 3109 images for training, 500 images for validation, and 1500 images for test. Compared with the aforementioned datasets, the difference is mainly reflected in two aspects. For one thing, it has much more diversities in scales, density and background. For another, it includes 351 negative samples (namely nobody scenes), which increase the variety of datasets. The quantitative results for NWPU-Crowd are listed in Table 6. It shows that the proposed GSANet scores 116.1 in MAE which ranks third, and 415.3 in RMSE which performs best among the trackers, respectively. Especially, compared with SCAR[52] which also adopts the attention mechanism, the proposed GSANet reduces the score of RMSE by 16.2%. Visualization of estimated maps with counting numbers is shown in Fig. 10. It proves that the proposed method performs well in the congested scenes with the accurate estimation.

Table 6 Experimental results on the NWPU-Crowd dataset.

Methods	MAE	RMSE
MCNN[14]	232.5	714.6
SANet[39]	190.6	491.4
A-CCNN[50]	176.5	520.6
ADMG[67]	152.8	907.3
RAZNet[68]	152.8	907.3
STANet[69]	122.6	468.3
PCC-Net[17]	112.3	457.0
SCAR[52]	**110.0**	495.3
GSANet (Ours)	116.1	**415.3**

Note: the best results are highlighted in **bold**.

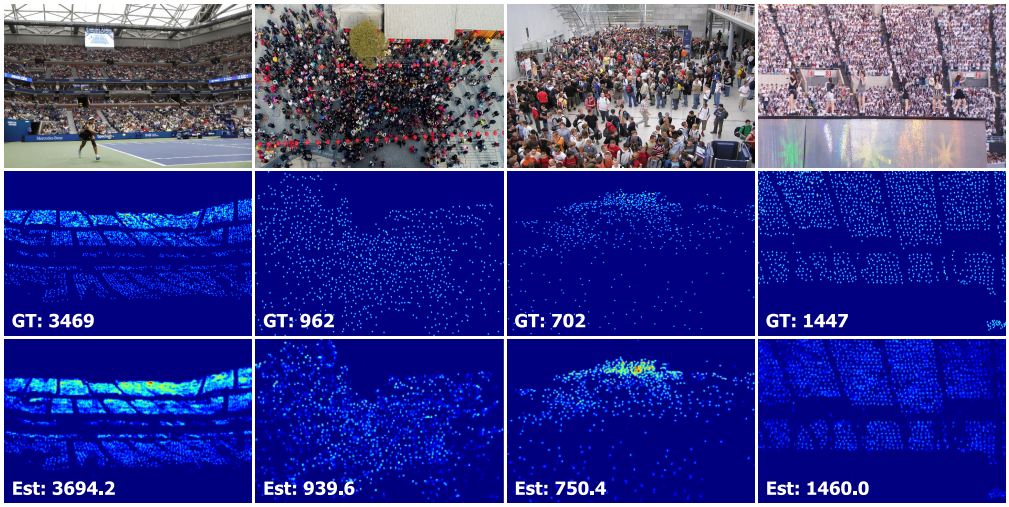

Fig. 10 The estimated density maps with counting number on exemplar images from the NWPU-Crowd dataset. Note: the first, second, and third rows represent the input image, the ground truth, and the estimated results, respectively.

4.3 *Ablation Study*

The effectiveness of critical components in GSANet are verified on ShanghaiTech_Part A dataset with different combinations. The counterparts are denoted as follows:

- "baseline" refers to the vanilla model without any component.
- "baseline + SA" denotes the baseline model with single SA unit.
- "baseline + CA" represents the baseline model with single CA unit.
- "baseline + CA-SA" denotes the baseline model with CA and SA units sequentially connected.
- "baseline + SA-CA" refers to the baseline model with SA and CA units sequentially connected.
- "baseline + SA || CA" represents the baseline model with CA and SA parallelly combined.
- "baseline + GS + SA || CA + AGG" represents adding the GS module and aggregation (AGG) module on the basis of the method denoted above.

The comparative results are shown in Table 7. It proves that all the critical components contribute to the substantial improvements of the baseline method. It can be observed that the CA unit performs better than SA unit in improving the accuracy. However, CA has a larger amount of calculation than SA. When the SA and CA units are sequentially connected (i.e., "baseline + CA_SA" method and "baseline + SA_CA" method), there is no effect in improving the accuracy, and it even degrades the accuracy. On the contrary, when the SA and CA are concatenated in parallel, i.e., "baseline + SA || CA" method, the scores of MAE and RMSE are reduced evidently compared with the baseline combined single SA or CA units. Compared with the "baseline + SA" and "baseline + CA" methods, the "baseline + SA || CA" method synergies the spatial and channel dimensional information, so that it can alleviate the estimation error in background regions. Therefore, the latter is better than the former. However, the accuracy is increased at the expense of complexity in terms of GFLOPs and parameters. To address this problem, the GS and AGG module are equipped and eventually led to the proposed GSANet method. It can be seen that, with the aids of GS and AGG modules, the GFLOPs and Params decline to 7.5 and 8.684 M, respectively. It exceeds all the other combined methods and even approaches to the baseline in efficiency. Furthermore, the final method scores 74.3 in MAE and 130.3 in RMSE, both outperforming other ensemble methods in accuracy. This can be attributed to the shuffle operation in AGG module, which facilitates the exchange of feature map of different channel dimensions.

Table 7 Ablation analysis on the key components in GSANet.

Methods	GFLOPs	Params (M)	MAE	RMSE
Baseline	7.498	8.674	88.4	146.5
Baseline + CA	7.505	9.461	80.0	135.0
Baseline + SA	7.500	8.685	80.5	136.2
Baseline + CA_SA	7.505	9.462	81.3	132.7
Baseline + SA_CA	7.504	9.462	81.0	148.6
Baseline + SA ‖ CA	7.505	9.461	76.7	131.5
Baseline + GS + SA ‖ CA + AGG	**7.500**	**8.684**	**74.3**	**130.3**

Note: the final results are highlighted in **bold**.

The qualitative comparisons of the different versions are shown in Fig. 11. Figure 11(a) is the exemplar image which is suffered from scale variation and background cluster. Figure 11(b) is the ground truth. Figure 11(c) indicates that the estimated results of the baseline deviate the ground truth to a large extent. The SA unit is helpful for the accurate location of heads, as depicted in red box of Fig. 11(d). The CA unit can alleviate the error estimation for background regions, as depicted in the green box of Fig. 11(e). The "baseline + CA-SA" makes the problem even worse, as shown in Fig. 11(f). The compound modes of "baseline + SA-CA" [Fig. 11(g)] and "baseline + SA ‖ CA" [Fig. 11(h)] boost the estimation accuracy, with the former being more

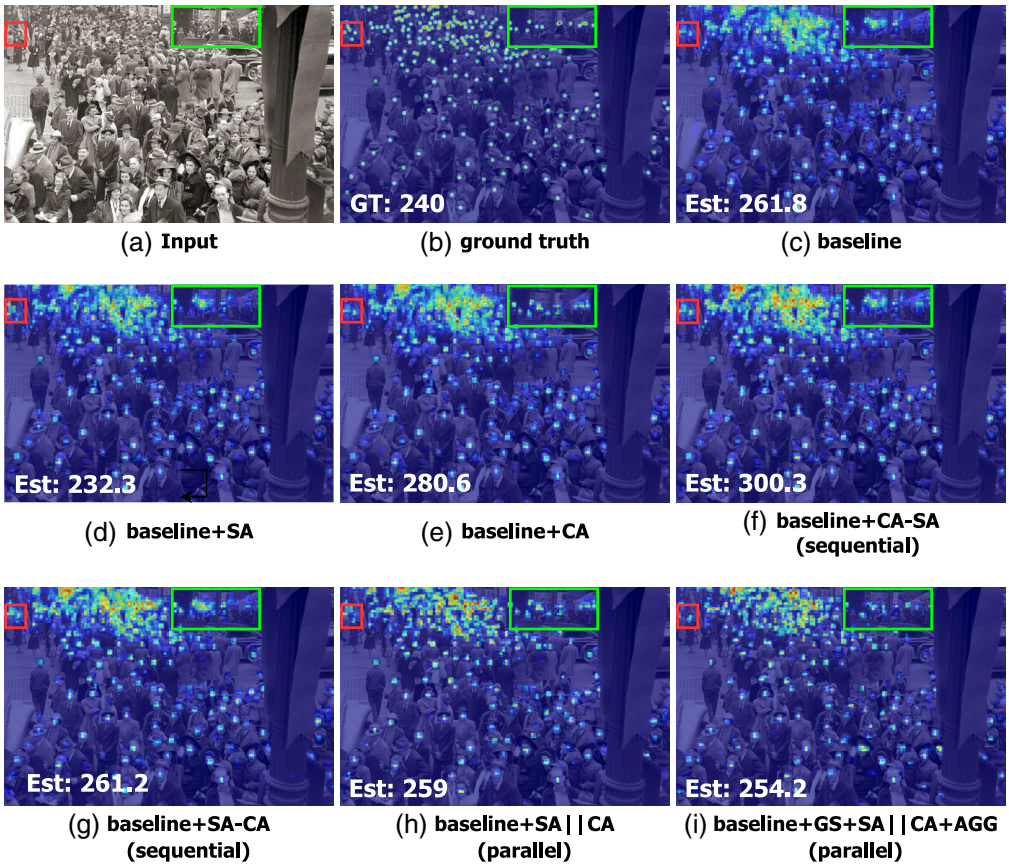

Fig. 11 The qualitative results of the baseline with different components.

effective. The final method [Fig. 11(i), GSANet] not only performs better in terms of pedestrian dispersion and estimated counts, but also it generates density maps that are closer to the ground truth within the green and red boxes.

5 Conclusion

GSANet is proposed for crowd counting under extremely high-density environment. The GSANet consists of three principal modules, namely GS module, dual-aware attention module, and aggregation module. The GS module reduces the calculation of network parameters, and makes the model more efficient. The attention module consists of a SA unit and a CA unit. The SA unit focuses on the spatial dependencies in the whole feature map to locate the heads accurately. The CA unit is built to explore the relations between channel maps and highlight the discriminative information in specific channels. The aggregation module enables information communication between different sub-features. Comparative experiments on five benchmark crowd datasets have proven the superiority of the proposed GSANet compared with the state-of-the-art competitors in accuracy and efficiency.

Acknowledgments

This work was supported in part 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. F. Ding et al., "Perceptual enhancement for autonomous vehicles: restoring visually degraded images for context prediction via adversarial training," *IEEE Trans. Intell. Transp. Syst.*, 1–12 (2021).
2. Z. Guo et al., "Hybrid intelligence-driven medical image recognition for remote patient diagnosis in internet of medical things," *IEEE J. Biomed. Health Inf.* (2021).
3. L. Tan et al., "Secure and resilient artificial intelligence of things: a honeynet approach for threat detection and situational awareness," *IEEE Consum. Electron. Mag.* **11**, 69–78 (2021).
4. D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 4031–4039 (2017).
5. V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.* **107**, 3–16 (2018).
6. M. Li et al., "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, pp. 1–4 (2008).
7. Z. L. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 604–618 (2010).
8. C. Zeng and H. Ma, "Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, pp. 2069–2072 (2010).
9. A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 1–7 (2008).
10. H. Idrees et al., "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 2547–2554 (2013).
11. H. Bai and S. Chan, "CNN-based single image crowd counting: network design, loss function and supervisory signal," ArXiv:abs/2012.15685 (2020).
12. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 886–893 (2005).

13. X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 32–39 (2009).
14. Y. Zhang et al., "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 589–597 (2016).
15. M. Hossain et al., "Crowd counting using scale-aware attention networks," in *Proc. IEEE Workshop Appl. of Comput. Vision (WACV)*, pp. 1280–1288 (2019).
16. J. Liu et al., "Decidenet: counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 5197–5206 (2018).
17. J. Gao, Q. Wang, and X. Li, "PCC Net: perspective crowd counting via spatial convolutional network," *IEEE Trans. Circuits Syst. Video Technol.* **30**, 3486–3498 (2020).
18. Z. Dong et al., "Scale-recursive network with point supervision for crowd scene analysis," *Neurocomputing* **384**, 314–324 (2020).
19. P. Wang et al., "Mobilecount: an efficient encoder-decoder framework for real-time crowd counting," *Neurocomputing* **407**, 292–299 (2020).
20. L. Liu et al., "Efficient crowd counting via structured knowledge transfer," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pp. 2645–2654 (2020).
21. I. Ahmed, M. Anisetti, and G. Jeon, "An IoT-based human detection system for complex industrial environment with deep learning architectures and transfer learning," *Int. J. Intell. Syst.* (2021).
22. P. Dollár et al., "Pedestrian detection: an evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 743–761 (2012).
23. F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 995–1008 (2016).
24. Q. Wang et al., "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 46–58 (2020).
25. A. C. Davies, J. Yin, and S. Velastin, "Crowd monitoring using image processing," *Electron. Commun. Eng. J.* **7**, 37–47 (1995).
26. A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)* (2017).
27. V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pp. 1324–1332 (2010).
28. J. Chen, W. Su, and Z. Wang, "Crowd counting with crowd attention convolutional neural network," *Neurocomputing* **382**, 210–220 (2020).
29. G. Gao et al., "Cnn-based density estimation and crowd counting: a survey," ArXiv:abs/2003.12783 (2020).
30. J. Zhang et al., "3D reconstruction for motion blurred images using deep learning-based intelligent systems," *Comput. Mater. Continua* **66**, 2087–2104 (2021).
31. Q. Zhang et al., "Graph neural networks-driven traffic forecasting for connected internet of vehicles," *IEEE Trans. Network Sci. Eng.* (2021).
32. B. Jan et al., "Deep learning in big data analytics: a comparative study," *Comput. Electr. Eng.* **75**, 275–287 (2019).
33. I. Ahmed et al., "A deep learning-based social distance monitoring framework for covid-19," *Sustain. Cities Soc.* **65**, 102571–102571 (2020).
34. V. A. Sindagi and V. M. Patel, "HA-CCN: hierarchical attention-based crowd counting network," *IEEE Trans. Image Process.* **29**, 323–335 (2020).
35. Y. Tian et al., "PaDNet: pan-density crowd counting," *IEEE Trans. Image Process.* **29**, 2714–2727 (2020).
36. Y. Li, X. Zhang, and D. Chen, "CSRNet: dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 1091–1100 (2018).
37. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 3431–3440 (2015).

38. Y. Zhang et al., "Attention to head locations for crowd counting," in *Proc. Int. Conf. Image and Graphics (ICIG)*, pp. 727–737 (2019).
39. X. Cao et al., "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 734–750 (2018).
40. S. Woo et al., "CBAM: convolutional block attention module," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 3–19 (2018).
41. X. Zhang et al., "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 6848–6856 (2018).
42. J. Gao, "c_3 framework: an open-source pytorch code for crowd counting," ArXiv:abs/1907.02724 (2019).
43. C. Zhang et al., "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 833–841 (2015).
44. V. A. Sindagi et al., "Learning to count in the crowd from limited labeled data," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 212–229 (2020).
45. V. Sindagi and V. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. IEEE Int. Conf. Adv. Video and Signal Based Surveill. (AVSS)*, pp. 1–6 (2017).
46. D. B. Sam and R. V. Babu, "Top-down feedback for crowd counting convolutional neural network," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, pp. 7323–7330 (2018).
47. Q. Wang et al., "Neuron linear transformation: modeling the domain shift for crowd counting," *IEEE Trans. Neural Networks and Learn. Syst.*, 1–13 (2021).
48. X. Shi et al., "A real-time deep network for crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 2328–2332 (2020).
49. L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proc. IEEE Workshop Appl. Comput. Vision (WACV)*, pp. 1113–1121 (2018).
50. S. A. Kasmani et al., "A-CCNN: adaptive CCNN for density estimation and crowd counting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 948–952 (2018).
51. Y. Lei et al., "Towards using count-level weak supervision for crowd counting," *Pattern Recognit.* **109**, 107616 (2021).
52. J. Gao, Q. Wang, and Y. Yuan, "Scar: spatial-/channel-wise attention regression networks for crowd counting," *Neurocomputing* **363**, 1–8 (2019).
53. D. B. Sam et al., "Locate, size, and count: accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 2739–2751 (2021).
54. Z. Yan et al., "Crowd counting via perspective-guided fractional-dilation convolution," *IEEE Trans. Multimedia* **24**, 2633–2647 (2021).
55. J. Cheng et al., "Decoupled two-stage crowd counting and beyond," *IEEE Trans. Image Process.* **30**, 2862–2875 (2021).
56. X. Jiang et al., "Attention scaling for crowd counting," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 4705–4714 (2020).
57. H. Idrees et al., "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 532–546 (2018).
58. L. Liu et al., "Denet: a universal network for counting crowd with varying densities and scales," *IEEE Trans. Multimedia* **23**, 1060–1068 (2021).
59. M. Hwan Oh, P. Olsen, and K. Ramamurthy, "Crowd counting with decomposed uncertainty," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, pp. 11799–11806 (2020).
60. D. Guo et al., "Dadnet: dilated-attention-deformable convnet for crowd counting," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pp. 1823–1832 (2019).
61. W. Zhai et al., "Da2net: a dual attention-aware network for robust crowd counting," *Multimedia Syst.*, 1–14 (2022).
62. Y. Wang et al., "Multi-scale dilated convolution of convolutional neural network for crowd counting," *Multimedia Tools Appl.* **79**, 1057–1073 (2019).
63. Q. Wang et al., "Density-aware curriculum learning for crowd counting," *IEEE Trans. Cybern.*, 1–13 (2020).

64. Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation," *IEEE Trans. Multimedia* **24**, 261–273 (2021).
65. W. Liu, M. Salzmann, and P. Fua, "Counting people by estimating people flows," *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).
66. Q. Wang et al., "Nwpu-crowd: a large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 2141–2149 (2021).
67. J. Wan and A. B. Chan, "Adaptive density map generation for crowd counting," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 1130–1139 (2019).
68. C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 1217–1226 (2019).
69. L. Wen et al., "Detection, tracking, and counting meets drones in crowds: a benchmark," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 7808–7817 (2021).

Wenzhe Zhai is pursuing his MS degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include image processing and pattern recognition.

Mingliang Gao received his PhD 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.

Marco Anisetti is an associate professor at the Universitä degli Studi di Milano, Italy. He received his PhD in computer science from the Universitä degli Studi di Milano in 2009. He is the winner of the GIRPR award for the best PhD 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.

Qilei Li is currently a PhD student at the School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom. He received his MS degree in signal and information processing from Sichuan University. His research interests are computer vision and deep learning.

Seunggil Jeon received his BS and MS degrees from Konkuk University, and his PhD from Hanyang University in 2008. He is currently a principal engineer at Samsung Electronics, Suwon, South Korea. He has extensive experience in industrial management of B5G wireless technologies. His current interests include IoT networks, wireless communications, and artificial intelligence in industry.

Jinfeng Pan received her PhD in signal and information processing from the University of Chinese Academy of Sciences, Xi'an, China, in 2016. She is currently an associate professor at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. Her main research interests include computer vision and deep learning.