Multimedia Systems (2023) 29:3027–3040
https://doi.org/10.1007/s00530-021-00877-4

SPECIAL ISSUE PAPER

DA^{2}Net: a dual attention-aware network for robust crowd counting

Wenzhe Zhai[1] · Qilei Li[2] · Ying Zhou[3] · Xuesong Li[1] · Jinfeng Pan[1] · Guofeng Zou[1] · Mingliang Gao[1]

Received: 24 July 2021 / Accepted: 6 December 2021 / Published online: 22 January 2022

Abstract
Crowd counting in congested scenes is a crucial yet challenging task in video surveillance and urban security system. The performance of crowd counting has been greatly boosted with the rapid development of deep learning. However, robust crowd counting in high-density environment with scale variations remains under-explored. To address this problem, we propose a dual attention-aware network (DA^{2}Net) for robust crowd counting in dense crowd scene with scale variations. Specifically, the DA^{2}Net consists of two modules, namely Spatial Attention (SA) module and Channel Attention (CA) module. The SA module focuses on the spatial dependencies in the whole feature map to locate the heads accurately. The CA module attempts to handle the relations between channel maps and highlights the discriminative information in specific channels. Thus, it alleviates the mistaken estimation for background regions. The interactions between SA module and CA module provide the synergy which facilitates the learning of discriminative features with a focus on the essential head region. Experimental results on five benchmark datasets, i.e., ShanghaiTech, UCF_CC_50, UCF-QNRF, WorldExpo'10, and NWPU, demonstrate that the DA^{2}Net can achieve the state-of-the-art performance on both accuracy and robustness.

Keywords Crowd counting · Density estimation · Attention mechanism · Convolutional neural network

1 Introduction

Crowd counting aims to predict the number of people or estimate the density maps for crowd scenarios. It has been a popular topic and arouse a great deal of interest in recent years, especially with the rapid growth of urban populations. Accurate crowd counting plays a crucial role in many real-world applications, such as video surveillance, crowd control, and public safety management [1, 2, 22, 46, 64]. Although crowd counting has drawn much attention, it is inherently challenging due to various degradation factors, e.g., scale variations, perspective distortion, serious occlusion, and non-uniform distribution.

To address these problems, a lot of efforts have been done in previous works which can be classified into three categories, namely detection-based methods, regression-based methods, and deep learning-based methods [46]. The detection-based methods attempt to estimate the number of people by detecting the body or head of each individual in the crowd. These methods lead to poor performance in highly dense crowd scene because of the poor detection performance in such scenarios. The regression-based methods dedicate to train regression models to directly map the visual features to the number of people. These methods ignore the spatial information as they are regressing on the global count. Benefiting from the powerful learning ability of deep convolutional neural networks (CNNs), CNN-based method have achieved commendable performance in crowd counting [1, 22, 46]. These approaches conduct crowd counting by learning density maps in an end-to-end manner.

Although the aforementioned methods have achieved great progress, they are still insufficient for real applications, especially with large-scale variations. The main challenge in crowd counting is the scale variations caused by the camera perspective distortion. Figure 1 illustrates the scale variation in crowd scenarios caused by perspective distortion. The large-scale variation will decrease the quality of estimated density maps, thus result in the error estimation for backgrounds.

✉ Mingliang Gao
mlgao@sdut.edu.cn

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

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

[3] School of Information Science and Engineering, Shandong University, Qingdao 266237, China

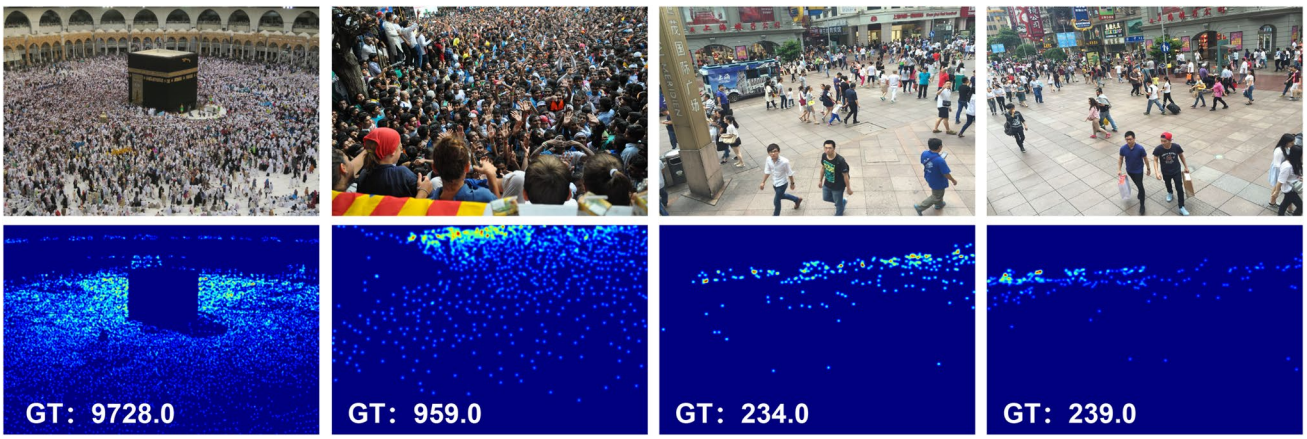

Fig. 1 Scale variations in the crowd scenes caused by perspective distortion. The first row depicts some representative samples, and the second row shows the corresponding ground truth density maps with counting numbers

In this paper, we address the problem of scale variation in dense crowd scenario and propose a robust crowd counting method named DA^{2}Net. The proposed method consists of two cascaded attention modules, i.e., Spatial Attention (SA) module and Channel Attention (CA) module. The SA module concentrates on the head region in the whole feature map. The CA module guides the network to focus on the relation between channel maps to eliminate the error estimation for background. Experimental results prove that the proposed DA^{2}Net achieves compelling performance on accuracy and robustness compared with the SOTA methods.

The paper is structured as follows. Section 2 presents an overview of the relevant prior works. Section 3 introduces the details of the proposed DA^{2}Net. Experiments are presented in Sect. 4. This work is concluded in Sect. 5.

2 Related work

In this section, we briefly review the relevant works from three aspects, i.e., detection-based methods, regression-based methods, and deep learning-based methods.

2.1 Detection-based methods

The early crowd counting approaches mainly adopt the detection-based schema. The detection-based methods mainly employ a sliding-window-like detector scanning the image, detect the body or head of each individual, and train a classifier to discriminate the positive samples. The crowd count is the number of positive samples outputted by the classifier, and the global detection scores are employed to estimate the crowd densities. Among these early approaches, Dollar [11] utilized a slide window detector over the image for person counting. Li et al. [24] estimated the number of people in the surveillance area by constructing detectors of head and shoulder. These methods locate peaks in foreground pixels to detect the heads or body parts, which depend on the exact foreground shape. The detection-based approaches work well in crowd scenarios with low-density, but they perform poorly in dense crowd scenes due to the issues of occlusion and background clutter [7].

2.2 Regression-based methods

Since the accuracy of detection-based method is unsatisfactory in highly congested scene, many regression-based methods are developed. The regression-based approaches aim at training regression models to map the visual feature maps to the number of people directly. The regression-based methods are proven to be feasible approaches in congested environments, as there is no need for explicit pedestrian segmentation and tracking. Davies et al. [9] first solved the problem of crowd statistics through regression methods. They extracted the underlying features of video frames to create a linear regression model of direct mapping of the overall feature information to the amount of people. Idress et al. [18] built a regression model to learn multi-features (i.e., head detection and SIFT [31]) of person in dense crowd images. Chen et al. [4] proposed an attribute-based cumulative regression approach to take into account the scalar variation of objects and achieve promising results on sparse data and imbalanced training data. Pham et al. [35] utilized a random forest regression to learn a non-linear mapping of feature and the number of people. Although the regression-based methods work well in high-density crowd scene, they ignore the crowd attention information and the spatial information, since the global count is declining.

2.3 Deep learning-based methods

Recently, thanks to powerful learning abilities, the deep learning-based methods have been introduced and achieved dominant performance in crowd counting. The main idea behind these methods is treating the task of crowd counting as a problem of estimating the density map, where the integral of the density map is used to estimate the number of persons in the corresponding region. Zhang et al. [56] introduced a deep CNN to address the across-scene counting problems. Marsden et al. [33] exploited the fully convolutional network (FCN) to handle the problem of crowd counting in highly dense scenarios. Zhang et al. [62] presented a multi-column CNN (MCNN) architecture to increase the receptive field to deal with the problem of scale variations. A contextual pyramid CNN (CP-CNN) [45] was built to fuse the global and local context information to generate high-quality density maps. Shen et al. [42] combined the cross-scale consistency pursuit loss and the adversarial loss together to tackle the problem of scale variations. Liu et al. [28] proposed a learnable spatial transform module with a region-wise refinement process to deal with both scale and rotation variations. Li et al. [25] built a deep CNN architecture for crowd counting by replacing the pooling operations with dilated kernels to obtain larger reception fields. Wang et al. [52] proposed a domain adaption framework for crowd counting by leveraging synthetic data.

Meanwhile, attention mechanism has become increasingly important in the field of deep learning, and it has been widely adopted in diverse vision domains [29, 51, 53, 59, 63]. The concept of attention is to focus selectively on a discrete aspect of information, both subjective and objective, while ignore other perceptible information [41]. The attention mechanism has also been adopted in crowd counting in recent years. Liu et al. [26] utilized an attention block termed QualityNet to capture the different importance weight of detection based map and regression based map by dynamically assessing the qualities of them for each pixel. Zhang et al. [61] exploited an attention model to generate a probability map to present higher probability scores in head regions. Kang et al. [21] utilized an image pyramid to deal with scale variations and proposed an across-scale attention map to softly select a suitable scale for each pixel. Hossain et al. [16] achieved the crowd counting by fusing a global scale attention module and a local scale attention module, among which the global scale attention module is to capture the overall density level of an image and the local scale attention module is to obtain the local scale information at different locations in an image. Zhang et al. [60] generated a score map using a multi-resolution attention model in which the response of the head position is higher than the non-head areas. Compared with the aforementioned attention methods, we devise a dual-aware attention model which consists of a spatial attention module and a channel attention module. The former module focuses on the spatial dependencies in the whole feature map and locates the heads accurately by channel-pool operations. The latter module handles the relations between channel maps and highlights the discriminative information in specific channels by a fast 1-dimension convolution and a sigmoid activation function. Then, the two modules are interacted to facilitate the learning of discriminative features with a focus on the essential head region.

3 Proposed method

3.1 Overview

Given an image I captured in congested scenes, the goal of crowd counting is to estimate the number of people by learning a discriminative model θ, which is capable of producing a faithful dense map $x : f_\theta(I) \rightarrow x$ to reflect the number of people. Considering the scale variations caused by perspective distortion, it is essential for the model θ to learning from the most significant regions while suppress the side effect brought by the disturbed patterns, e.g., scale variation. To this aim, we propose the dual attention-aware network (DA^{2}Net) which explores the mutual reinforcement of spatial attention (SA) and channel attention (CA). The flowchart of the proposed DA^{2}Net is shown in Fig. 2.

Supposing there is a feature map $F \in \mathbb{R}^{C\times H\times W}$ extracted from the corresponding input image I, the SA module generates a 2-dimension spatial attention map $\mathrm{O}_{\mathrm{sa}}(F) \in \mathbb{R}^{H\times W}$ to reflect the importance in spatial space regarding vital regions (e.g., head). The spatial-enhanced feature map F_{s} is obtained by

$$F_{\mathrm{s}} = \mathrm{O}_{\mathrm{sa}}(F) \otimes F, \tag{1}$$

where $\mathrm{O}_{\mathrm{sa}}(\cdot)$ is the function of SA module, and $\otimes$ denotes the element-wise multiplication. The spatial-enhanced feature map F_{s} serves as the bridge between SA and CA modules, which is further refined by exploring the complementary channel-wise importance with the proposed CA,

$$F_{\mathrm{cs}} = \mathrm{O}_{\mathrm{ca}}(F_{\mathrm{s}}) \oplus F_{\mathrm{s}}. \tag{2}$$

Where $\mathrm{O}_{\mathrm{ca}}(\cdot)$ is the function of CA module, and $\oplus$ denotes the element-wise sum. $F_{\mathrm{cs}} \in \mathbb{R}^{C\times H\times W}$ is the enhanced feature map, which is more discriminative than the original counterpart F in both spatial and channel spaces. Based on F_{cs}, we infer the density map to achieve the counting of people.

3.2 Spatial attention module

Due to the perspective changes of crowd scenarios, the distribution of head region in both global and local view has a

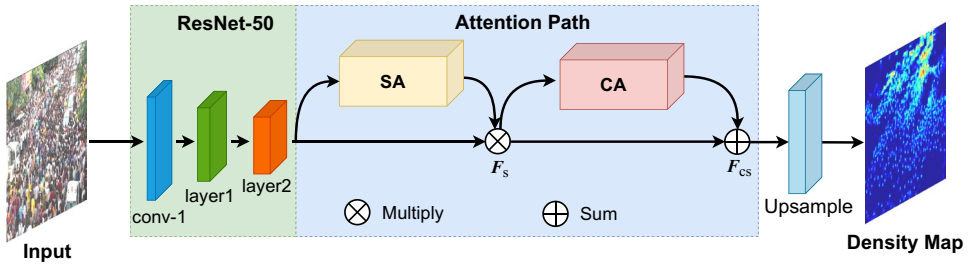

Fig. 2 The flowchart of the DA^{2}Net for crowd counting. The proposed approach consists of two modules, i.e., SA and CA in a cascaded pattern, to encode the local range and global contextual feature maps, respectively. It concatenates these two types of feature maps and then produces a 1-channel predicted density map via upsample operation

certain degree of regularity. From the global point of view, it is consistent with a gradual change in density. For instance, as shown in Fig. 1, the density maps are increasing from proximal to distal due to the perspective distortion. In terms of local image patches with high density, their textures and patterns are similar.

To encode the aforementioned two observations, we build the SA module to focus on the head region, and guarantees the location of the head accurately. The architecture of the SA module is shown in Fig. 3. It is denoted as follows,

$$\mathrm{O}_{\mathrm{sa}} = \mathrm{Sigmoid}(\mathrm{C}_7([\mathrm{MaxPool}(F)];\mathrm{AvgPool}(F)]), \tag{3}$$

where the symbol C_7 represents a convolution layer with the kernel size of 7×7. MaxPool$(\cdot)$ and AvgPool$(\cdot)$ denote the max pooling and average pooling on channel dimensions, respectively.

3.3 Channel attention module

The aforementioned SA module attempts to encode the dependencies on a spatial dimension, which can locate the head position accurately. However, it will result in error estimation for the background due to the resemblances between foreground and background region texture. To address this problem, we design the complementary CA module as shown in Fig. 4. The CA module is formulated as follows.

$$\begin{aligned} g_c &= \frac{1}{WH}\sum_{i=1,j=1}^{W,H} F_c(i,j), \\ \mathrm{O}_{ca} &= \mathrm{Sigmoid}(C1D_k(g_c)), \end{aligned} \tag{4}$$

where g_c represents the channel-aware global average pooling (GAP) which obtains the aggregated features of background region. The CA module generates channel weights by performing a fast 1-dimension convolution ($C1D$). $F_c = \left[F_c^{i,j}\right]_{H\times W} \in \mathbb{R}^{H\times W}, c \in \{1, 2, \ldots, C\}$ represents the feature map corresponding to each channel.

3.4 Loss function

We adopt the pixel-wise Mean Square Error (MSE) to measure the difference between the estimated density map and the ground truth. Given an image I_i, the learnable parameter θ of DA^{2}Net is optimized as follows,

$$\mathrm{loss} = \frac{1}{N}\sum_{i=1}^{N} \left\| f_\theta(I_i) - Y_i \right\|_2^2, \tag{5}$$

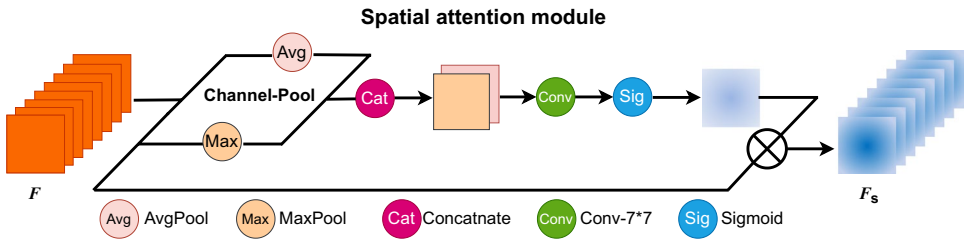

Fig. 3 The architecture of the spatial attention (SA) module. The channel-pool operation employs the max pooling and average pooling on channel dimensions to get different feature vectors which are fused through concatenation. Then, the spatial attention map is generated by a convolution operation and a sigmoid activation function

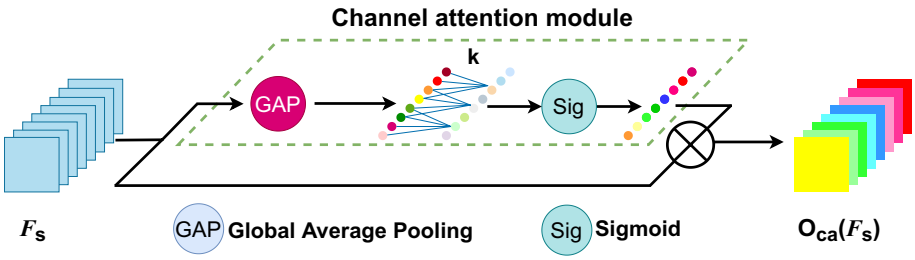

Fig. 4 The architecture of the channel attention (CA) module. It takes the high level feature map F_s as input and employs the global average pooling (GAP) operation to generate a channel-wise weight vector $g_c \in \mathbb{R}^C$. Then, the channel weights are generated by performing a fast 1-dimension convolution and a sigmoid activation function. Finally, the channel attention map is generated by the product of channel weights and the input F_s

where N is the batch size of input mini-batch, $f_\theta(I_i)$ denotes the estimated results, and Y_i is the corresponding ground truth density map.

3.5 Implementation details

3.5.1 Ground truth density map

The Ground truth density map $H(x)$ is generated with the geometry-adaptive Gaussian kernel G_σ and convolving with a delta function. The formula is defined as follows,

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x), \tag{6}$$

where N is the number of head annotations, x corresponds to a pixel in the image, and x_i represents the coordinates of the head annotation. The delta function $\delta(x - x_i)$ is utilized to depict a head. When the delta function is equal to 1, there is a head in this pixel.

3.5.2 Training details

In this work, the experimental training and evaluation are implemented on two paralleled NVIDIA RTX2080S GPU using PyTorch framework [12]. All the images and the corresponding density maps are resized to 576×768. The Adam optimizer is adopted. The learning rate is initialized at 10^{-5} and reduced to 0.995 times per epoch to minimize the training loss.

4 Experiments

Experiments on five benchmark datasets, i.e., ShanghaiTech [62], UCF_CC_50 [18], UCF-QNRF [19], WorldExpo'10 [56], and NWPU [47], are conducted to compare the DA^2 Net with other SOTA methods.

4.1 Evaluation metrics

Following the general evaluation principles [25, 44, 56], we adopt the Mean Absolute Error (MAE) and Mean Square Error (MSE) as the evaluation metrics. The definitions of the MAE and MSE are indicated in formula (7) and (8), respectively.

$$\text{MAE} = \frac{1}{N} \sum |y_i - \hat{y}_i|, \tag{7}$$

$$\text{MSE} = \sqrt{\frac{1}{N} \sum |y_i - \hat{y}_i|^2}, \tag{8}$$

where N denotes the number of images to be tested, $\hat{y}_i$ is the estimated count value of the i-th testing image, and y is the corresponding ground truth density map. Generally, MAE and MSE indicate the accuracy and robustness of the crowd estimation, respectively [6, 55].

4.2 Performance on ShanghaiTech dataset

The ShanghaiTech dataset [62] contains 1,198 images with 330,165 annotations. It is divided into two parts, i.e., Part_A and Part_B. The former part consists of 482 images (300 images for training and 182 images for testing) which were randomly collected from the Internet. The later part includes 716 images (400 images for training and 316 images for testing) which were taken from the urban areas in Shanghai city.

Table 1 Experimental results on the ShanghaiTech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Zhang et al. [56]	181.8	277.7	32.0	49.8
Marsden et al. [33]	126.5	173.5	23.8	33.1
MCNN [62]	110.2	173.2	26.4	41.3
TDF-CNN [37]	97.5	145.1	20.7	32.8
Switching-CNN [40]	90.4	135.0	21.1	30.1
DecideNet [26]	–	–	20.8	29.4
BSAD [17]	90.4	135.0	20.2	35.6
TDF-CNN [37]	97.5	145.1	20.7	32.8
C-CNN [43]	88.1	141.7	14.9	22.1
SaCNN [57]	86.8	139.2	20.7	32.8
A-CCNN [23]	85.4	124.6	11.0	19.0
PCC-Net [13]	**73.5**	124.0	19.2	31.5
DNCL [58]	**73.5**	**112.3**	18.7	26.0
DA^{2}Net(ours)	74.1	128.4	**7.9**	**13.2**

Best results are marked in bold

The quantitative results for ShanghaiTech dataset are depicted in Table 1. It shows that the proposed DA^{2}Net obtains the score of 73.5 and 112.3 in MAE and MSE on Part_A, and 7.9 and 13.2 on Part_B. Especially, the DA^{2}Net ranks the first in both MAE and MSE on Part_B. Although the DA^{2}Net does not acquire the first place on Part_A, it is comparable to the most popular SOTA methods. Figure 5 illustrates some representative results on Shanghai Part_A, and Part_B. It shows that the proposed method is robust to the interference from scale variation. The estimated crowd density map can depict the densities of different regions, and the estimated counting numbers are approximate to the ground truth value.

4.3 Performance on UCF_CC_50 dataset

The UCF_CC_50 dataset [18] includes 50 images collected from highly-crowded scenes. The crowd counts within it ranging from 94 to 4543, and the average is 1280. Following the general principle [18], we adopt 5-fold cross-validation strategy in evaluation.

Experimental results on the UCF_CC_50 dataset are depicted in Table 2. It indicates that the proposed DA^{2}Net achieves a score of 169.5 in MAE and a score of 237 in MSE, both ranking first among the compared methods. This reveals the exceptional performance on accuracy and robustness in highly dense crowd scenes. Especially, compared with SCAR [14] and ASNet [20] which also adopt the attention mechanism in crowd density estimation, the proposed DA^{2}Net reduces the score of MAE by 34.56% and 3%, MSE by 36.63% and 5.8%, respectively. In order to highlight the

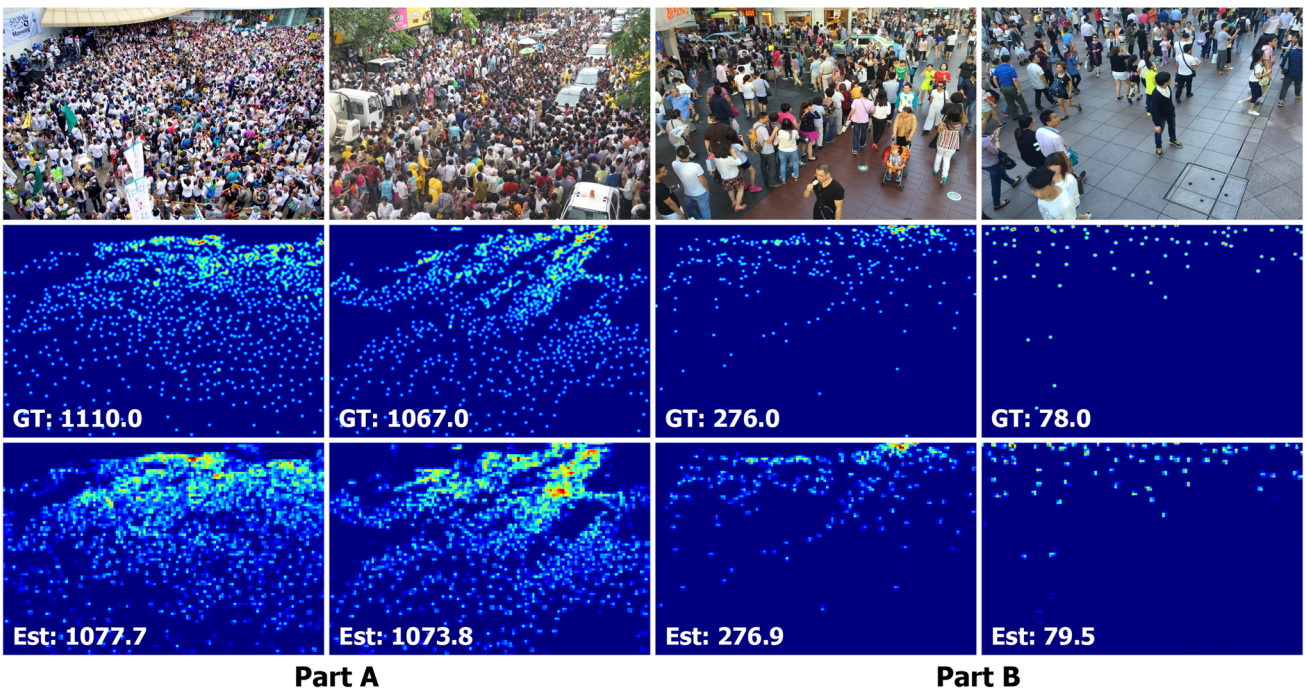

Fig. 5 Exemplar images of the ShanghaiTech_Part_A and ShanghaiTech_Part_B datasets(the first row), the ground truth (the second row), and the estimated results (the third row)

Table 2 Experimental results on the UCF_CC_50 dataset

Methods	MAE	MSE
Idrees et al. [18]	419.5	541.6
Zhang et al. [56]	467.0	498.5
MCNN [62]	377.6	509.1
Switching-CNN [40]	318.1	439.2
CMTL [44]	322.8	397.9
SaCNN [57]	314.9	424.8
CP-CNN [45]	295.8	320.9
DR-ResNet [10]	307.4	421.6
CSRNet [25]	266.1	397.5
ic-CNN [36]	260.9	365.5
SCAR [14]	259.0	374.0
ASNet [20]	174.8	251.6
D2CNet [8]	182.1	254.9
DA^{2}Net (ours)	**169.5**	**237.0**

Best results are marked in bold

effectiveness of the proposed method intuitively, we perform subjective evaluations with three representative methods, i.e., MCNN [62], CRSNet [25], and SCAR [14]. Specifically, the MCNN [62] was also proposed to solve the problem of head-scale variations caused by perspective effect. The CSRNet [25] is a competitive method which expands the receptive field by the dilated convolutions. The SCAR [14] takes into account the large-range pixel-wise contextual information and utilizes attention mechanism to improve the estimated accuracy. The subjective results are compared in Fig. 6. One can see that compared with the three competitors, the crowd density maps and counting numbers estimated by the proposed method are closer to the ground truth even in high-density scenarios.

4.4 Performance on UCF-QNRF dataset

The UCF-QNRF dataset [19] consists of 1535 challenging images (1201 images for training and 334 images for testing) with 1,251,642 manual annotations. The minimum, maximum, and mean counts are 49, 12,865, and 815.4 respectively. Also, the images in this dataset are captured in the wild, making it more realistic as well as difficult.

Table 3 shows the comparative results on the UCF-QNRF [19] dataset. It can be observed that the proposed DA^{2}Net achieves the lowest MAE of 111.7 and the third lowest MSE of 204.3. Figure 7 illustrates the qualitative results for sample images from the UCF-QNRF datasets. It depicts that the DA^{2}Net has superior robustness against both the density and scale variation.

4.5 Performance on WorldExpo'10 dataset

WorldExpo'10 dataset [56] is a large data-driven cross-scene benchmark dataset for crowd counting. It contains 3380 frames in 103 scenes in the training set and 600 labelled frames from the remaining 5 scenes in the testing set. Following the general setting [49], we set an ROI in which the number of people to be counted.

The performance of our model against other SOTA methods are reported in Table 4. It indicates that the proposed method outperforms all the other SOTA methods, except for the Scene 4. The average value of MAE is 7.03, and it ranks the first place among the compared method which verifies the efficiency of the proposed method for diverse scenes. Exemplar qualitative results of the WorldExpo'10 dataset are shown in Fig. 8. The experimental results indicate that the proposed method is robust and effective in both dense scenes and sparse scenes.

4.6 Performance on NWPU dataset

The NWPU dataset [47] is currently the largest dataset for crowd counting and localization. It includes 5109 images with 2,133,375 head annotations. Compared with other datasets, it has many challenge factors such as containing negative samples, high-resolution and large appearance changes.

The subjective evaluation of the DA^{2}Net model with SOTA methods are reported in Table 5. It demonstrates that the proposed DA^{2}Net gains the results of 102.6 and 378.5 in MAE and MSE, which are both the best results compared with other SOTA method. Compared with another attention-based method, i.e., SCAR [14], the proposed DA^{2}Net reduces the score of MSE by 23.58% which indicates that the DA^{2}Net is more robust than SCAR [14] for the large-scale congested crowd counting. Some exemplar images with the corresponding density maps are shown in Fig. 9. One can see that the DA^{2}Net generates a density map which is highly closed to the corresponding ground truth.

4.7 Cross-dataset analysis

To demonstrate the generalization ability of the proposed model, cross-dataset analysis is conducted in this section. We take the ShanghaiTech Part_A as the training set, and the UCF_50 and UCF-QNRF datasets as the test sets, respectively. Three representative methods, i.e., MCNN [62], CRSNet [25], and SCAR [14] are used as the competitors. The comparative results are reported in Table 6. It shows that compared with the three competitors, the proposed method attains the best results in both MAE and MSE, which verifies the competitive generalization ability of the proposed method.

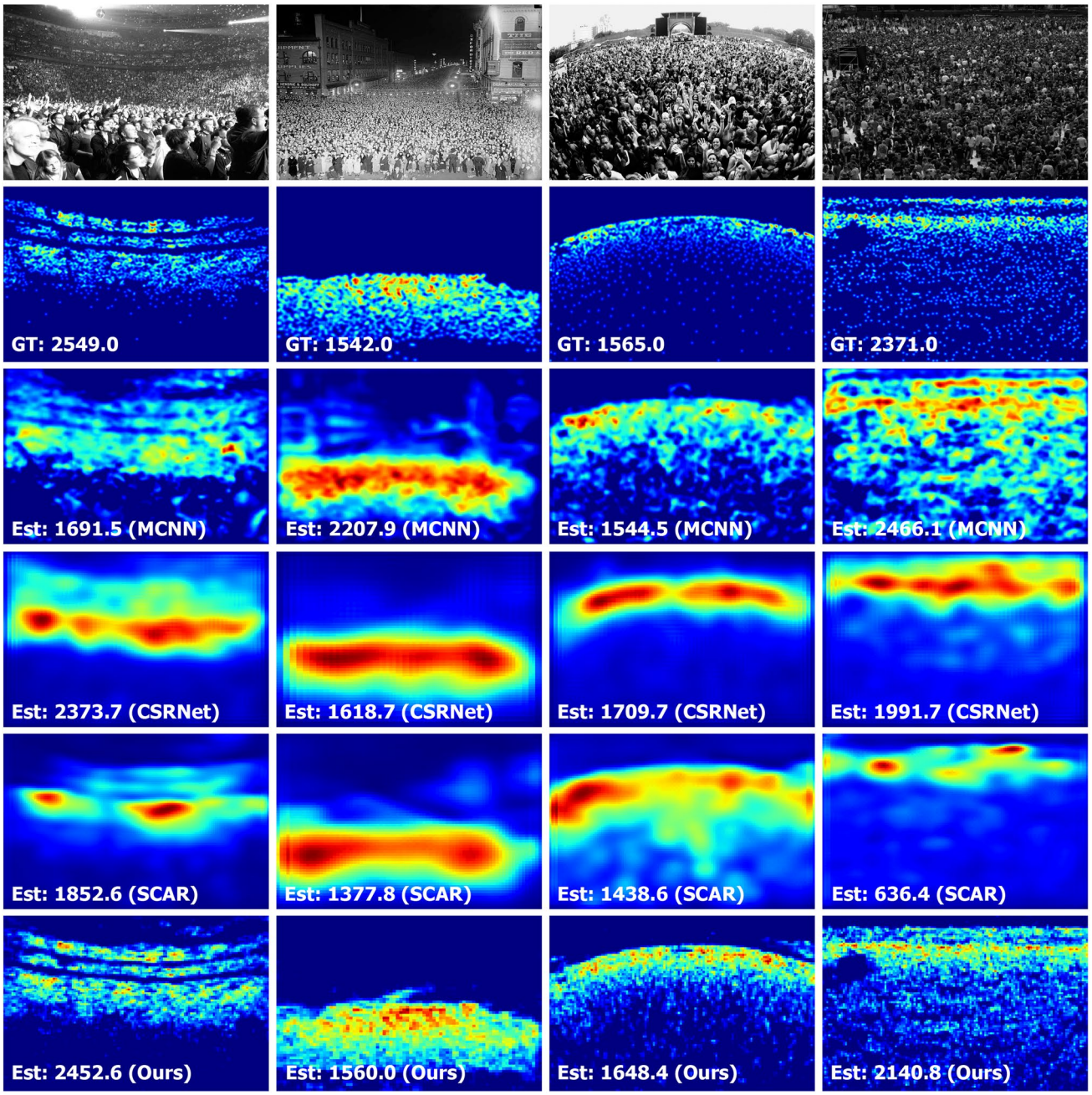

Fig. 6 Exemplar images of the UCF_50 dataset(the first row), the ground truth (the second row), the estimated results of MCNN (the third row), the estimated results of CSRNet (the fourth row), the estimated results of SCAR (the fifth row), and the estimated results of DA^{2}Net (the sixth row)

4.8 Ablation study

The effectiveness of critical components in DA^{2}Net are demonstrated by designing several counterparts with different combinations. Counterparts are denoted as follows. (1) 'baseline' refers to the vanilla model that does not adopt the SA module and CA module. (2) 'baseline+SA' denotes the baseline model with solely SA module. (3) 'baseline+CA' denotes the baseline model by solely combing CA module. (4) 'baseline+SA || CA' represents the SA and CA are parallelly combined into the baseline. (5) 'baseline+CA-SA' denotes the CA and SA modules are sequentially connected following the CA-first, SA-second order. (6) 'baseline+SA-CA' also sequentially connected following the SA-first,

Table 3 Experimental results on the UCF-QNRF dataset

Methods	MAE	MSE
Zhang et al. [56]	467.0	498.5
Idress et al. [18]	315.0	508.0
MCNN [62]	277.0	509.1
CMTL [44]	252.0	514.0
SCAR [14]	264.8	418.3
PCCNet [13]	148.7	247.3
Switching-CNN [40]	228.0	445.0
CRSNet [25]	129.0	209.0
DENet [27]	121.0	205.0
LSC-CNN [38]	120.5	218.2
DADNet [15]	113.2	189.4
DUBNet [34]	116.0	**178.0**
DA^{2}Net (ours)	**111.7**	204.3

Best results are marked in bold

CA-second order as opposite to 'baseline+CA-SA' and 'baseline+SA || CA'.

The overall quantitative performance is shown in Table 7. It indicates that the two critical components, i.e., SA module and CA module, contribute to the substantial improvement of the baseline method in terms of both MAE and MSE. The SA module is more effective that CA module in improving the accuracy and robustness. The final DA^{2}Net boost the baseline significantly by 13.2% and 11.4% in terms of MAE and MSE, respectively.

The qualitative comparison of the baseline with different components is shown in Fig. 10. Although the number of people in this exemplar image is small (70), it is still challenging as it suffered from scale variations and background cluster, simultaneously, as shown in the red and green box in Fig. 10a. It shows that baseline method is suffered from the background cluster and scale variations. The estimated number and the density map deviate the ground truth to a large extent, as depicted in Fig. 10c. The SA module guarantees the accurate location of heads, as depicted in the green box in Fig. 10d. The CA module can alleviate the error estimation for background regions, as depicted in the red box in Fig. 10e. The 'baseline+CA-SA' mode in Fig. 10g makes the problem worse. Both the compound modes of 'baseline+SA || CA' (Fig. 10f and 'baseline+SA-CA' (Fig. 10h) boost the estimation accuracy, with the latter being more effective.

4.9 Failure cases

The proposed DA^{2}Net outperforms most the SOTA methods by a large margin. However, there are some failure cases under the extremely challenging scenarios, as shown in Fig. 11 shows. Specifically, When the crowd scenarios are captured in highly low-light environment, the estimated density map and the counting results deviate the ground

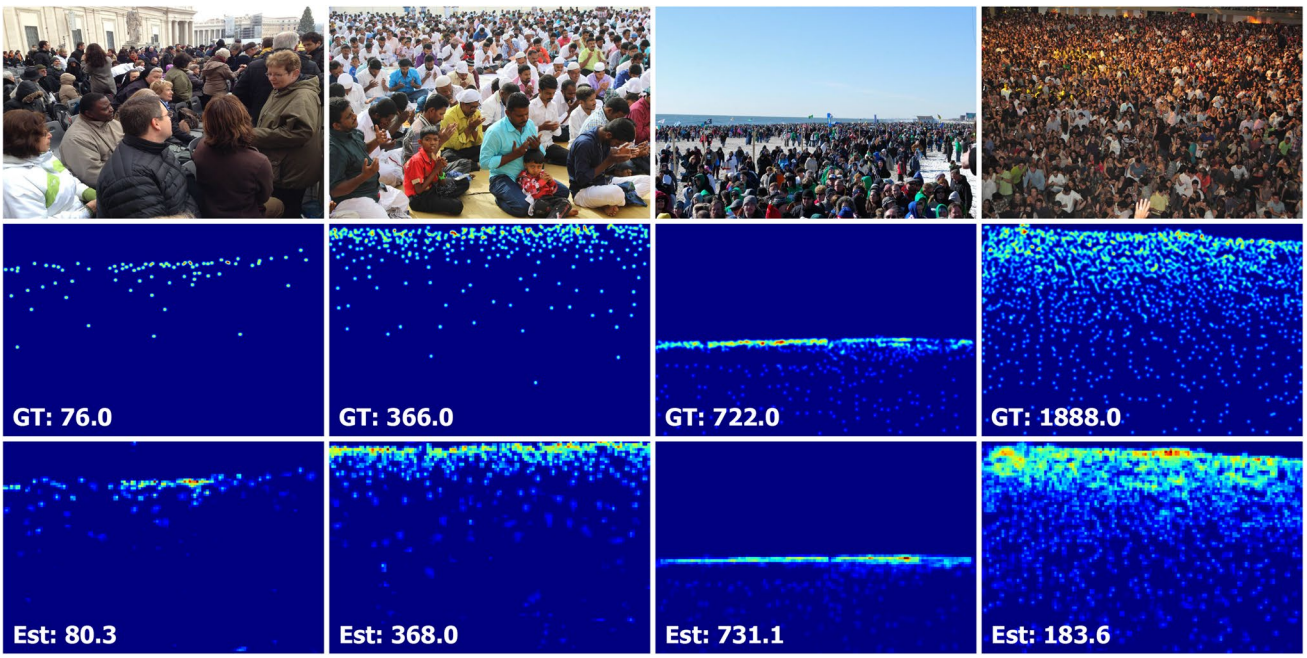

Fig. 7 Exemplar images of the UCF-QNRF dataset (the first row), the ground truth (the second row), and the estimated results (the third row)

Table 4 Experimental results on the WorldExpo'10 dataset

Methods	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	MAE(Avg.)
LBP+RR [5]	13.6	58.9	37.1	21.8	23.4	31.9
Zhang et al. [56]	9.8	14.1	14.3	22.4	3.7	12.9
MCNN [62]	3.4	20.6	12.9	13.0	8.1	11.6
MSCNN [52]	7.8	15.4	14.9	11.8	5.8	11.7
ConvLSTM-nt [54]	8.6	16.9	14.6	15.4	4.0	11.9
SCAR [14]	1.9	13.8	9.6	29.8	3.9	11.8
TDF-CNN [37]	2.7	23.4	10.7	17.6	3.3	11.5
IG-CNN [39]	2.6	16.1	10.15	20.2	7.6	11.3
PSCC+DCL [50]	1.8	16.2	9.2	25.0	2.8	11.0
ic-CNN [36]	17.0	12.3	9.2	**8.1**	4.7	10.3
CSRNet [25]	2.9	11.5	8.6	16.6	3.4	8.6
SANet [3]	2.6	13.2	9.0	13.3	3.0	8.2
DRSAN [28]	2.6	11.8	10.3	10.4	3.7	7.76
DA^{2}Net (ours)	**1.45**	**11.8**	**7.95**	11.6	**2.35**	**7.03**

Best results are marked in bold

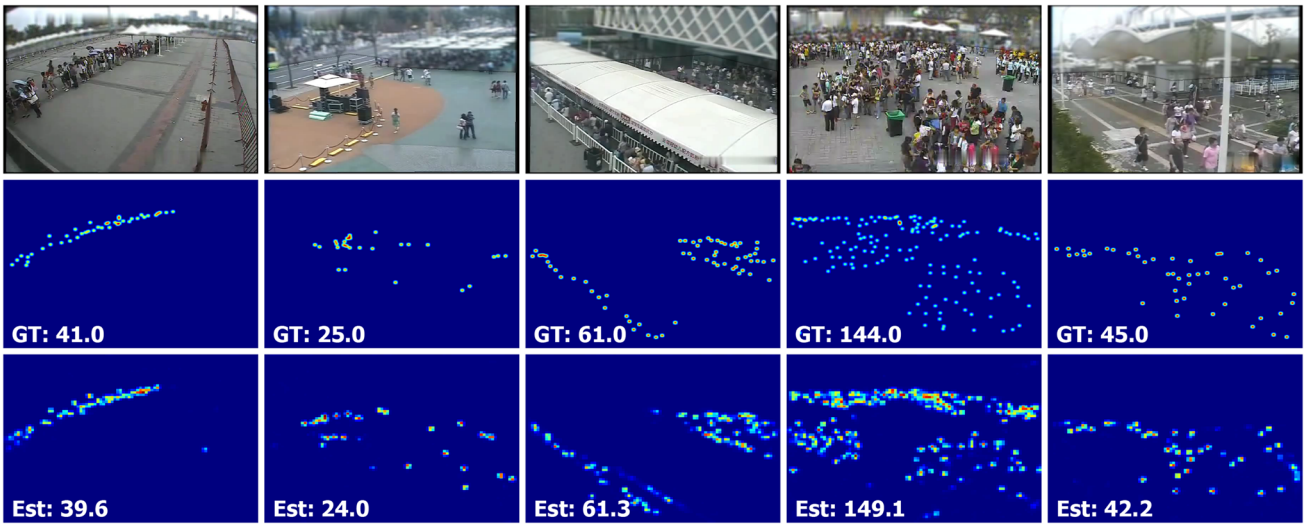

Fig. 8 Exemplar images of the WorldExpo'10 dataset (the first row), the ground truth (the second row), and the corresponding estimated density map (the third row). Each column represents a typical scene

Table 5 Experimental results on the NWPU dataset

Methods	MAE	MSE
MCNN [62]	232.5	714.6
CRSNet [25]	121.3	378.8
SANet [3]	190.6	491.4
PCC-Net [13]	112.3	457.0
CAN [30]	106.3	386.5
SCAR [14]	110.0	495.3
BL [32]	105.4	454.2
SFCN [48]	105.7	424.1
DA^{2}Net(ours)	**102.6**	**378.5**

Best results are marked in bold

truth to a large degree. Crowd counting in low-light environment is a difficult problem as the feature maps in head region are very close to the background area. Future work is expected on reliable crowd feature map extraction in low-light environment.

5 Conclusion

The scale variation in crowd scenario is a primary degradation factor in crowd counting, which degrades the accuracy of the crowd estimation. To address this problem, we

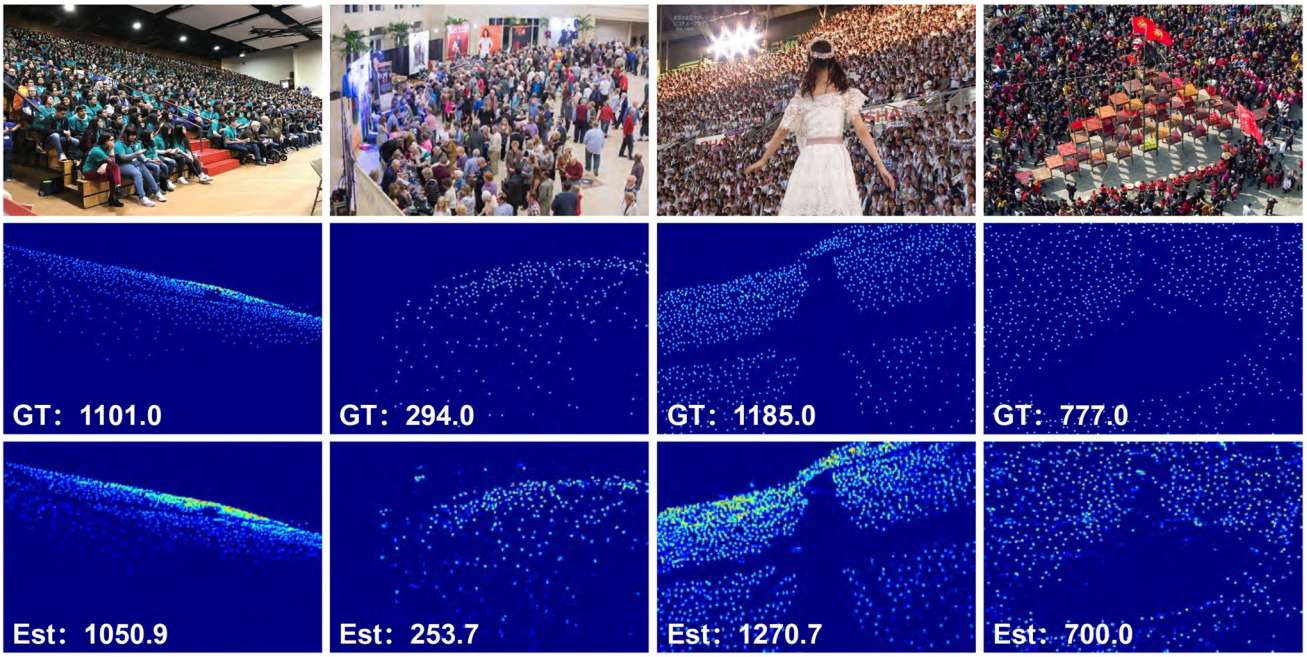

Fig. 9 Exemplar images of the NWPU dataset (the first row), the ground truth (the second row), and the estimated results (the third row)

Table 6 Comparative results on the cross-data testing

Methods	Source dataset	Target dataset	MAE	MSE
UCF_50 Dataset				
MCNN [62]	UCF_50	UCF_50	377.6	509.1
CSRNet [25]	UCF_50	UCF_50	266.1	397.5
SCAR [14]	UCF_50	UCF_50	259.0	374.0
DA^{2}Net(ours)	UCF_50	UCF_50	**169.5**	**237.0**
UCF_50 Cross-Dataset				
MCNN [62]	ShanghaiTech Part_A	UCF_50	496.5	709.5
CSRNet [25]	ShanghaiTech Part_A	UCF_50	409.9	604.6
SCAR [14]	ShanghaiTech Part_A	UCF_50	470.6	686.4
DA^{2}Net(ours)	ShanghaiTech Part_A	UCF_50	**379.0**	**567.7**
UCF-QNRF Dataset				
MCNN [62]	UCF-QNRF	UCF-QNRF	277.0	509.1
CSRNet [25]	UCF-QNRF	UCF-QNRF	129.0	209.0
SCAR [14]	UCF-QNRF	UCF-QNRF	264.8	418.3
DA^{2}Net(ours)	UCF-QNRF	UCF-QNRF	**111.7**	**204.3**
UCF-QNRF Cross-Dataset				
MCNN [62]	ShanghaiTech Part_A	UCF-QNRF	340.3	571.9
CSRNet [25]	ShanghaiTech Part_A	UCF-QNRF	193.1	375.2
SCAR [14]	ShanghaiTech Part_A	UCF-QNRF	262.9	499.8
DA^{2}Net(ours)	ShanghaiTech Part_A	UCF-QNRF	**172.5**	**330.3**

Best results are marked in bold

Table 7 Ablation analysis of the key components in DA^{2}Net on ShanghaiTech Part B dataset

Methods	MAE	MSE
Baseline	9.1	14.9
Baseline+SA	8.6	13.9
Baseline+CA	8.9	14.0
Baseline+SA ‖ CA	8.6	13.3
Baseline+CA-SA	9.4	16.1
Baseline+SA-CA	**7.9**	**13.2**

Best results are marked in bold

propose a dual attention-aware network (DA^{2}Net) which consists of a spatial attention module and a channel attention module. The former module guarantees the accurate location of heads, while the latter module alleviates the error estimation for background regions. These two modules highlight the crucial information in spatial and channel spaces in a mutual-promotion manner. Comprehensive experiments on five benchmark datasets prove that the DA^{2}Net achieves compelling performance on accuracy and robustness compared with the SOTA methods.

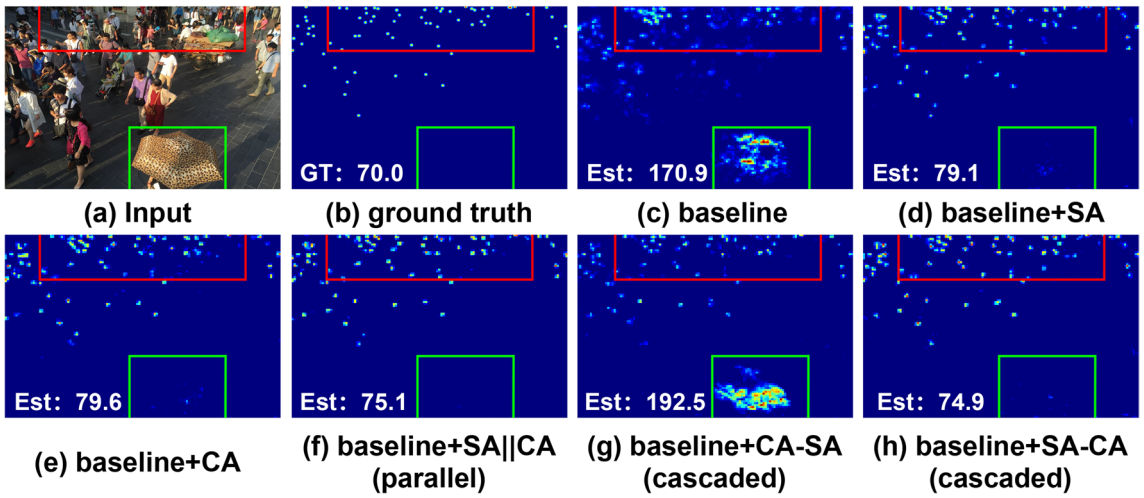

Fig. 10 The qualitative comparison of the baseline with different components. The red box depicts the scale variations cased by perspective, and the blue box illustrates the background cluster

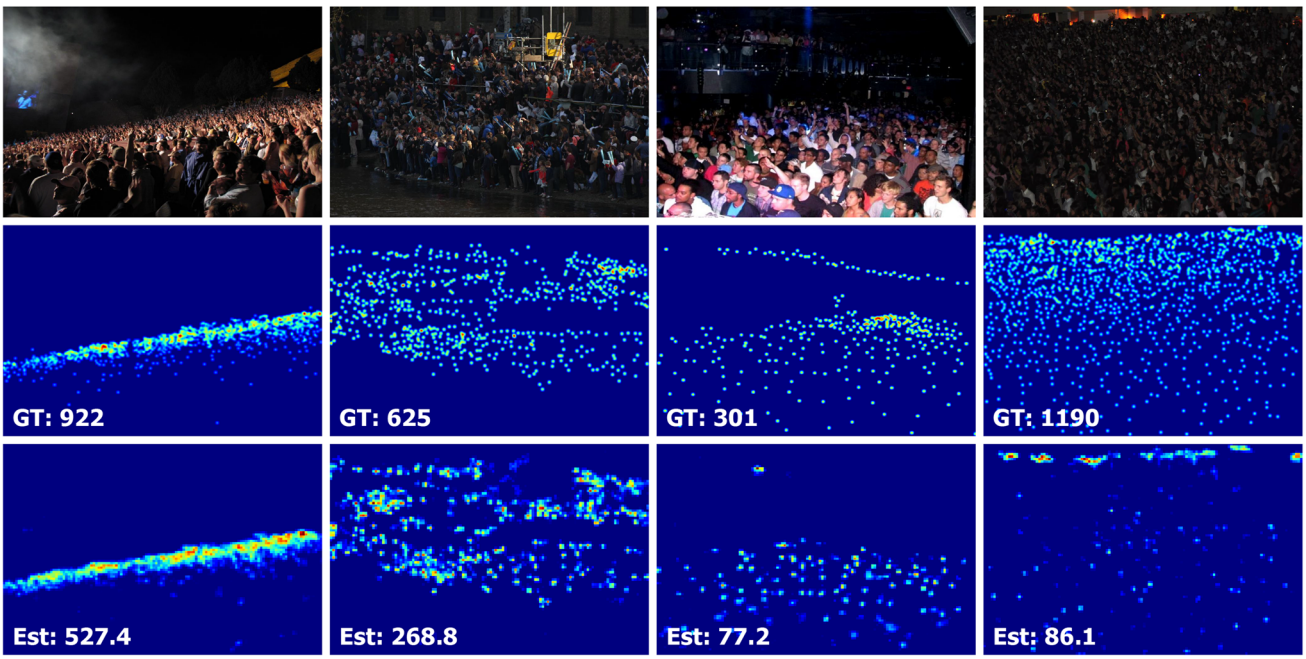

Fig. 11 The failure cases. The first, second and third rows are the exemplar images, ground truth and estimated results, respectively

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 61801272), the National Natural Science Foundation of Shandong Province (Nos.ZR2021QD041 and ZR2020MF127), and Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (No. 2019JZZY010119).

References

1. Bai, H., Chan, S.: Cnn-based single image crowd counting: Network design, loss function and supervisory signal. ArXiv arXiv: abs/2012.15685 (2020)
2. Ben, X., Ren, Y., Zhang, J., Wang, S.J., Kpalma, K., Meng, W., Liu, Y.: Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1–1 (2021)

3. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
4. 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), pp. 2467–2474 (2013)
5. Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localised crowd counting. In: Proceedings of the British Machine Vision Conference (BMVC), p. 3 (2012)
6. Chen, X., Bin, Y., Sang, N., Gao, C.: Scale pyramid network for crowd counting. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV), pp. 1941–1950 (2019)
7. Chen, X., Yan, H., Li, T., Xu, J., Zhu, F.: Adversarial scale-adaptive neural network for crowd counting. Neurocomputing **450**, 14–24 (2021)
8. Cheng, J., Xiong, H., Cao, Z., Lu, H.: Decoupled two-stage crowd counting and beyond. IEEE Trans Image Process **30**, 2862–2875 (2021)
9. Davies, A.C., Yin, J., Velastin, S.: Crowd monitoring using image processing. Electron Commun Eng J **7**, 37–47 (1995)
10. Ding, X., Lin, Z., He, F., Wang, Y., Huang, Y.: A deeply-recursive convolutional network for crowd counting. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1942–1946 (2018)
11. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 743–761 (2012)
12. Gao, J., Lin, W., Zhao, B., Wang, D., Gao, C., Wen, J.: c^3 framework: An open-source pytorch code for crowd counting. ArXiv arXiv:abs/1907.02724 (2019)
13. 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)
14. Gao, J., Wang, Q., Yuan, Y.: Scar: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing **363**, 1–8 (2019)
15. Guo, D., Li, K., Zha, Z., Wang, M.: Dadnet: Dilated-attention-deformable convnet for crowd counting. In: Proceedings of the ACM International Conference on Multimedia (ACM MM) (2019)
16. Hossain, M., Hosseinzadeh, M., Chanda, O., Wang, Y.: Crowd counting using scale-aware attention networks. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV), pp. 1280–1288 (2019)
17. Huang, S., Li, X., Zhang, Z., Wu, F., Gao, S., Ji, R., Han, J.: Body structure aware deep crowd counting. IEEE Trans Image Process **27**, 1049–1059 (2018)
18. 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), pp. 2547–2554 (2013)
19. 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), pp. 532–546 (2018)
20. Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., Pang, Y.: Attention scaling for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4705–4714 (2020)
21. Kang, D., Chan, A.B.: Crowd counting by adaptively fusing predictions from an image pyramid. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
22. Kang, D., Ma, Z., Chan, A.B.: Beyond counting: comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. IEEE Trans Circuits Syst Video Technol **29**, 1408–1422 (2019)
23. 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), pp. 948–952 (2018)
24. Li, M., Zhang, Z., Huang, K., Tan, T.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–4 (2008)
25. 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), pp. 1091–1100 (2018)
26. 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), pp. 5197–5206 (2018)
27. 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 Multimedia **23**, 1060–1068 (2021)
28. Liu, L., Wang, H., Li, G., Ouyang, W., Lin, L.: Crowd counting using deep recurrent spatial-aware network. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 849–855 (2018)
29. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), pp. 843–851 (2018)
30. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5094–5103 (2019)
31. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1150–1157 (1999)
32. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 6141–6150 (2019)
33. Marsden, M., McGuinness, K., Little, S., O'Connor, N.: Fully convolutional crowd counting on highly congested scenes. In: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), pp. 27–33 (2017)
34. Mini-hwan O., Olsen, P., Ramamurthy, K.: Crowd counting with decomposed uncertainty. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 11799–11806 (2020)
35. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2017)
36. Ranjan, V., Le, H.M., Hoai, M.: Iterative crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 278–293 (2018)
37. Sam, D.B., Babu, R.V.: Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2018)
38. Sam, D.B., Peri, S., Sundararaman, M.N., Kamath, A., Babu, R.V.: Locate, size and count: accurately resolving people in dense crowds via detection. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 2739–2751 (2021)

39. 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), pp. 3618–3626 (2018)
40. 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), pp. 4031–4039 (2017)
41. de Santana Correia, A., Colombini, E.: Attention, please! a survey of neural attention models in deep learning. ArXiv arXiv:abs/2103.16775 (2021)
42. Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., Yang, X.: Crowd counting via adversarial cross-scale consistency pursuit. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5245–5254 (2018)
43. 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), pp. 2328–2332 (2020)
44. 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), pp. 1–6 (2017)
45. Sindagi, V., Patel, V.: Generating high-quality crowd density maps using contextual pyramid cnns. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1879–1888 (2017)
46. 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)
47. 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)
48. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8190–8199 (2019)
49. Wang, Q., Han, T., Gao, J., Yuan, Y.: Neuron linear transformation: Modeling the domain shift for crowd counting. IEEE transactions on neural networks and learning systems **PP** (2021)
50. Wang, Q., Lin, W., Gao, J., Li, X.: Density-aware curriculum learning for crowd counting. IEEE Transactions on Cybernetics pp. 1–13 (2020)
51. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539 (2020)
52. Wang, Y., Hu, S., Wang, G., Chen, C., Pan, Z.: Multi-scale dilated convolution of convolutional neural network for crowd counting. Multimed Tool Appl **79**, 1057–1073 (2019)
53. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
54. Xiong, F., Shi, X., Yeung, D.: Spatiotemporal modeling for crowd counting in videos. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 5161–5169 (2017)
55. Yang, B., Cao, J., Wang, N., Zhang, Y., Zou, L.: Counting challenging crowds robustly using a multi-column multi-task convolutional neural network. Signal Process. Image Commun. **64**, 118–129 (2018)
56. 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), pp. 833–841 (2015)
57. 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), pp. 1113–1121 (2018)
58. 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)
59. Zhang, X., Liu, M., Yin, J., Ren, Z., Nie, L.: Question tagging via graph-guided ranking. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), pp. 1–23 (2022)
60. Zhang, Y., Zhou, C., Chang, F., Kot, A.: Multi-resolution attention convolutional neural network for crowd counting. Neurocomputing **329**, 144–152 (2019)
61. Zhang, Y., Zhou, C., Chang, F., Kot, A.C.: Attention to head locations for crowd counting. In: Proceedings of the International Conference on Image and Graphics (ICIG), pp. 727–737 (2019)
62. 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), pp. 589–597 (2016)
63. Zhao, Y., Nie, W., Liu, A.A., Gao, Z., Su, Y.: Svhan: Sequential view based hierarchical attention network for 3d shape recognition. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), pp. 2130–2138 (2021)
64. Zitouni, M.S., Bhaskar, H., Dias, J., Al-Mualla, M.: Advances and trends in visual crowd analysis: a systematic survey and evaluation of crowd modelling techniques. Neurocomputing **186**, 139–159 (2016)