# CDADNet: Context-guided dense attentional dilated network for crowd counting

Aichun Zhu, Guoxiu Duan, Xiaomei Zhu, Lu Zhao, Yaoying Huang, Gang Hua, Hichem Snoussi

# CDADNet: Context-guided dense attentional dilated network for crowd counting

Aichun Zhu[a,b], Guoxiu Duan[a], Xiaomei Zhu[a], Lu Zhao[a],
Yaoying Huang[a], Gang Hua[b], Hichem Snoussi[c]

a School of Computer Science and Technology, Nanjing Tech University, Nanjing, China
b School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China c ICD - LM2S, Université de Technologie de Troyes, France

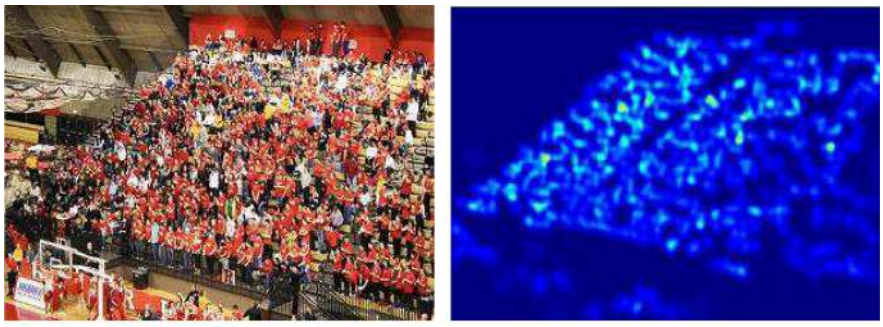
Crowd counting is a conspicuous task in computer vision owing to scale variations, perspective distortions, and complex backgrounds. Existing research usually adopts the dilated convolution network to enlarge the receptive fields to solve the problem of scale variations. However, these methods easily bring background information into the large receptive fields to generate poor quality density maps. To address this problem, we propose a novel backbone called Context-guided Dense Attentional Dilated Network (CDADNet). CDADNet contains three components: an attentional module, a context-guided module and a dense attentional dilated module. The attentional module is used to provide attention maps which can remove background information, while the context-guided module is proposed to extract multi-scale contextual information. Moreover, the dense attentional dilated module aims to generate high-granularity density maps and the cascaded strategy is used to preserve information from changing scales. To verify the feasibility of our method, we compare it to the existing approaches on five crowd counting datasets (ShanghaiTech (Part_A and Part_B), WorldEXPO'10, UCSD, UCF_CC_50). The comparison results demonstrate that CDADNet is effective and robust for various scenes.

## 1. Introduction

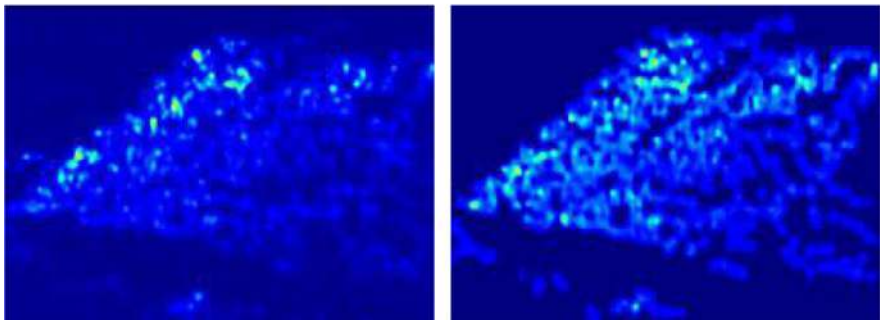
In recent years, with the application of crowd counting in flow analysis and safety assurance, great progress has been made in image-based population density detection. Many researchers have been devoted to exploring the algorithms and models to achieve better performance of crowd counting. However, crowd counting is still a remarkable work, with large-scale variations, perspective distortions, and complex backgrounds.

Recently, many researchers attempt to solve the above problems by applying convolutional neural networks. [1–3] proposed multi-column/branch dilated networks to cope with the problem of massive scale variations. In [4], a DSNet is proposed to capture a large range of scales and achieve superb performance. Yan et al. [5] introduced PGC, as an insertable module, to address scale variation in a single image. In the PGCNet model, a perspective estimation branch was trained to produce a perspective map guiding PGC to achieve good performance on density estimation. In [6], to handle the inconsistency caused by pixel-wise independence of density map, RANet was proposed, in which LSA and GSA were used to capture local and global pixel-wise interdependencies respectively. As for the problem of perspective distortion, it has been solved by combining density maps extracted from different resolution image blocks [7] or feature maps obtained with multi-scale contextual information [1,8]. In the Ref. [9], Shi et al. proposed PACNN architecture to address the issue of perspective distortion through predicting perspective maps and integrating the person scale information into density regression. Shi et al. [10] designed D-ConvNet to deal with the over-fitting phenomenon on a single image, in which the model applied new learning strategies containing NCL and 'divide and conquer' to achieve better trade-offs among the bias–variance–covariance and product more generalizable features. Different from the study of generating estimated density maps, the Ref. [11] mainly concentrated on density map generation. They used Refiner subnetwork to iteratively refine the original density map to produce a more reasonable density map in the training stage. In other ways, Zhao et al. [12] formulated extra three heterogeneous loss including geometric, semantic and numeric loss for backbone CNN to obtain more robust crowd representations and high-quality estimated density maps. Meanwhile, the attention mechanism [13,14] is a common solution to the problem of scenes with a complex background. However, existing methods usually focused on the problem of scale variations or perspective distortion. This will cause these models to bring background information easily, which will adversely affect the representation of scale or background features, and

* Corresponding author at: School of Computer Science and Technology, Nanjing Tech University, Nanjing, China.
*E-mail address:* aichun.zhu@njtech.edu.cn (A. Zhu).



(a) Original image (b) Ground truth



(c) CSRNet without attention (d) CSRNet with attention

**Fig. 1.** In the first row, (a) is an image from the ShanghaiTech dataset, (b) shows the ground truth density map. In the second row, (c) shows the generated density map by CSRNet without attention information, and (d) shows the generated density map by CSRNet with an attention module.

only work well on some specific crowd counting datasets. As shown in Fig. 1, a standard dilated network CSRNet [3] is modeled with an attentional module that could significantly weaken the influence of the background information and improve its ability to represent the scale and context information.

Large-scale variations and complex backgrounds will cause great errors in crowd counting. In order to address these problems and obtain accurate counting results, we propose a novel architecture called Context-guided Dense Attentional Dilated Network (CDADNet). This model can generate high-quality density maps by extracting multi-scale information and erasing background information. The CDADNet contains three components: an attention module, a context-guided module, and a dense attentional dilated module. As shown in Fig. 2, the first 13 layers of the Vgg-16 network are used as the frontend network of the proposed model. In the attention module, a Dense Dilated Block (DDB) is used to generate attention maps to remove background information and pay close attention to the foreground crowd information. Then, the context-guided module takes the output of Vgg-16, followed by the average pooling to extract multi-scale contextual information. Moreover, the dense attentional dilated module is proposed to generate high-quality density maps. Four cascaded Dense Attentional Dilated Blocks (DADBs) are added to the dense attentional dilated module and take the multi-scale contextual features as an input. And each DADB is constructed of a DDB multiplying the generated attention maps. In this network, the context-guided module extracts multi-scale contextual information to handle scale variations, and attention mechanism is firstly added to the density expansion network to effectively solve the problems of no-people background. Finally, the experimental results indicate that the CDADNet achieves excellent performance on the crowd counting tasks for extremely dense scenes and relatively sparse scenes. What is also worth mentioning is that, compared with the existing methods, our method achieves the best performance, not only in the counting accuracy but also in the stability of the model.

In summary, this paper mainly makes the following contributions.

(1) We design a context-guided module that concatenates multi-scale contextual features to provide rich contextual information for Dense Attentional Dilated Blocks(DADBs). The output density map was obtained by inputting contextual feature maps into four stacked DADBs. The whole network, i.e. CDADNet, can be trained end-to-end and can generate high-granularity density maps.

(2) In addition to the general Euclidean loss, we additionally introduce the cross-entropy loss and the adaptive density-level loss(ADLoss). The cross-entropy loss was applied to optimize the generation of attention map. The ADLoss can better address the estimation error caused by uneven density level.

(3) The results were extensively tested on five challenging crowd counting datasets (ShanghaiTech (Part_A and Part_B), WorldEXPO'10, UCSD, UCF_CC_50), and the method consistently outperform other outstanding methods.

Other parts of this article are described as follows: In Section 2, we briefly review related work of CNN-based and attention-based models for crowd counting. Section 3 mainly introduces our proposed network of this paper, namely the Context-guided Dense Attentional Dilated Network (CDADNet). The experiment including training details, ablation study, and results comparison is presented in Section 4. The conclusion of this paper is presented in Section 5.

## 2. Related work

### *2.1. CNN-based models for crowd counting*

Crowd counting is a complicated problem in terms of the difficulty in foreground extraction and overcrowd between the crowd. Previous studies have confirmed that the CNN-based method is effective for crowd counting [2,7,15], most of which use a multi-column framework. A Multi-column Convolutional Neural Network (MCNN) was described in [1] to process images of any size and predict crowd density with three networks, each network with different convolutional kernel sizes. Considering the heavy calculations of this method, Sam et al. proposed a regressor for density map prediction by several CNNs, and then the optimal CNN regressor was selected for each input image, the best one of which was used as the final result (Switch-CNN) [2].

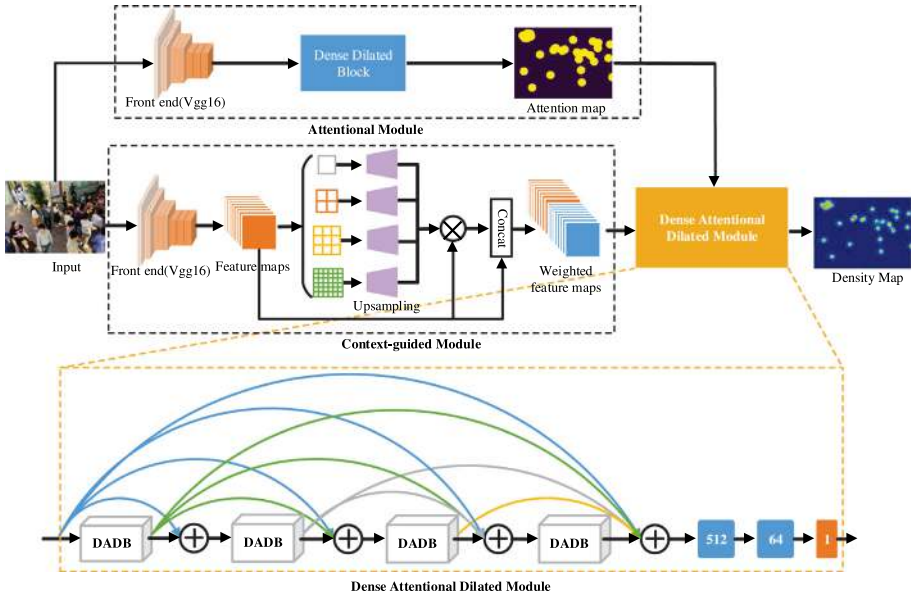



**Fig. 2. The architecture of CDADNet.** CDADNet contains three components: the attentional module, the context-guided module, and the dense attentional dilated module. With a given input image, Vgg-16 serves as the front-end network. Then the attentional module is used to generate attention maps, while the context-guided module performs average pooling operation on the output of Vgg-16 to obtain features of different scales. Finally, the dense attentional dilated module is constructed of four Dense Attentional Dilated Blocks (DADBs) which generates high-quality density maps. It should be noted that each DADB is built by a Dense Dilated Block (DDB) with an attention module.

Moreover, Sindagi et al. designed a novel method called CP-CNN which merges global and local context information to produce superior crowd density maps and number estimation [15]. Furthermore, Onoro-Rubio et al. exploited a Hydra CNN to learn a multi-scale model by using columns as pyramid levels on image patches [7]. Despite much success achieved in crowd counting, these methods still have several problems when used; for example, a large amount of parameters would lead to increasing difficulty in training and consume much of time. To overcome these shortcomings, researchers have attempted to adopt multi-scale single-column architecture [3,16,17]. Li et al. proposed to use dilated convolution to replace some pooling layers in the CNN which enlarges the receptive field without increasing parameters and calculations (CSRNet) [3]. Then, to combine the features of multi-scale information and dilated convolution, Yang et al. proposed an ASPP [18] sampling the given input on different sampling rates in parallel, which is equivalent to capturing the context of the image at multiple scales. A similar structure is used in this paper to cover large receptive fields and choose different dilation rates so that it can achieve more context information than the methods mentioned above.

### *2.2. Attention-based models for crowd counting*

Recently, attention models have been widely used in various areas of deep learning, such as image classification [19], semantic segmentation [20], object detection and classification [21], and crowd counting [22]. Many papers related to crowd counting cover attention mechanism. These models generated weights for spatial distribution on feature maps, and then try to make the network learning pay attention to different areas of objects selectively, which can help to select the most relevant information for visual analysis. For semantic segmentation, it is necessary to embed multi-scale information. Chen et al. [20] exploited the attention model to measure the importance of different scale features after generating multi-resolution inputs. On top of that, Liu et al. proposed a DANet which uses a self-attention mechanism to capture rich semantic information and improve the discrimination of feature representation [23]. More recently, inspired by Wang et al. [24], more attention modules added in the residual attention network can linearly improve the performance of classification of the network, and attention models can be extracted from feature maps of different depths. The CACrowdGAN [22] was proposed to abate the effect of complex background and generate density maps by the cascaded discriminator. Besides, attention can be applied in most of the current deep networks to achieve end-to-end training results. Because of the existence of the residual structure, the network can be easily extended to hundreds of layers, and using this strategy can significantly reduce the amount of calculation. To our knowledge, the proposed CDADNet for the first time has incorporated the attention mechanism into a dense dilated network to improve the quality of density maps.

## 3. Proposed method

The architecture of the proposed Context-guided Dense Attentional Dilated Network (CDADNet) is illustrated in Fig. 2. It consists of three components: context-guided module, attentional module, and dense attentional dilated module. The context-guided module is used to learn the weights for the context-aware features. The attentional module is designed to extract attentional information from the input images. In addition, the dense attentional dilated module is proposed to enlarge the attentional receptive fields and provide high-quality density maps. In the following sections, we will detail the architecture of the CDADNet.

### *3.1. Attentional module*

Visual attention is an essential mechanism of the human brain for understanding scenes effectively. Therefore, we aim to make the network focus on crowd regions in the input image using the attention mechanism. In our work, a two-category classification network is used to classify an input image into foreground regions and background regions. In this subsection, considering that the previous methods easily bring the background information into the large receptive fields to generate poor quality density maps. A novel attentional module is proposed to weaken background information for generating high-density maps. The attentional module is built by the first 13 layers from VGG-16 and a Dense Dilated Block (DDB). The DDB is composed of three dilated convolutional layers with a dilation rate of 1, 2, 3. As shown in Fig. 3, the dense dilated block is the basic component of the Dense Attentional Dilated Block (DADB). A DADB is constructed of a DDB multiplying a

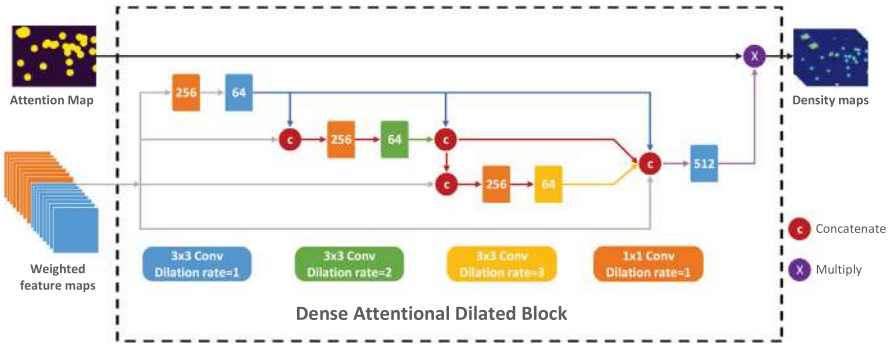



**Fig. 3. The architecture of the Dense Attentional Dilated Block (DADB).** A DADB takes the weighted features from the context-guided module as the input. It is constructed by a Dense Dilated Block (DDB) multiplying a generated attention map. The DDB is composed of three dilated convolutional layers with dilation rate of 1, 2, 3.

generated attention map. Let $y[j]$ represents output and $I[j]$ represents input, a DDB usually can be formulated as follows:

$$y[j] = \sum_{k=1}^{K} I[j + d \cdot k] \cdot w[k] \tag{1}$$

where $d$ is expressed as dilation rate, $k$ represents the size of the filter. $w[k]$ denotes the parameters of the $k$th filter. It presents that a large dilation rate has a large receptive field.

### 3.2. Context-guided module

Here we address the problem of large-scale variations and perspective distortions of crowd counting. To better represent multi-scale information of the input image, a context-guided module is incorporated into the proposed model. As shown in Fig. 2, the front-end is built by the first 13 layers of a pre-trained Vgg-16 network, and the output of Vgg-16 serves as the input of the context-guided module. Thus, the context-guided module aims to construct multi-scale contextual information from Vgg-16 features. When given an input image I, its output features can be defined as:

$$f_o = F_{vgg}(I) \tag{2}$$

As discussed above, the limitation of the standard Vgg-16 model encodes the same receptive fields. To overcome this limitation, we utilize multi-scale contextual features by performing average pooling and a $1 \times 1$ convolutional layer. Then these multi-scale features are concatenated together, serving as the input of DADB. Here we represent these multi-scale contextual features as:

$$f_s = U_{bi}(F_s(P_{ave}(f_o, s), v_s)) \tag{3}$$

where, for each scale $s$, $P_{ave}$ averages $f_o$ into $k(s) \times k(s)$ blocks. $F_s$ is a convolutional network with a kernel size of 1 to connect different channels of the contextual features. In the experiment, we use four different scales with corresponding block sizes $v_s \in 1, 2, 3, 6$.

### 3.3. Dense attentional dilated module

Although the combination of multi-scale information and dilation convolution features with different dilation rates can increase the receptive field, the larger receptive field usually brings more background information. Hence, we proposed a dense attentional dilated module that contains several Dense Attentional Dilated Blocks (DADBs). These blocks contain three dilated convolutional layers with dilation rates of 1, 2, 3. The setting can reduce the loss of pixel information due to a series of interprime dilations. Specifically, the output of the previous DADB serves as the input to the following DADB, as shown in Figs. 2 and 3. And to prevent the network from getting too long and losing information, we add an attention generator behind each block to better integrate shallow and deep information. The proposed dense attentional dilated module contains four DADBs, and each DADB is constructed by a DDB multiplying a generated attention map. This cascaded structure enables the proposed network to have a larger receptive field without extra background information. In this way, the proposed model can generate higher quality density maps.

### 3.4. Loss function

We introduce the loss function of the proposed CDADNet. Here, the cross-entropy loss is introduced to the attention module, and the Euclidean loss is used for crowd counting. The attention loss can be expressed as:

$$L_{att} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (M_{ij} log(\hat{M}_{ij}) + (1 - M_{ij}) log(1 - \hat{M}_{ij})) \tag{4}$$

where $W$ and $H$ denote the width and length of the image. $M_{ij}$ and $\hat{M}_{ij}$ represent the attentional mask from the ground truth and the generated density map. Moreover, the distinction between the ground truth and the generated density map is measured by the Euclidean distance. Thus, the Euclidean loss function is defined as:

$$L(\theta) = \frac{1}{2N} \sum_{j=1}^{N} \| D(I_j; \theta) - D_j^{GT} \|_2^2 \tag{5}$$

where $N$ is the batch_size and $D(I_j; \theta)$ is the generated density map by CDADNet with the weighted parameter $\theta$. $D_j^{GT}$ is a ground truth of input image $I_j$. In the work of training, this loss neglects the effect of different densities levels. The distribution of the images containing low-density, medium-density and high-density regions is quite different. Therefore, the standard loss function will cause the model training to be biased towards estimating the density. To cope with the estimate errors caused by unbalanced density levels, we presented Adaptive Density-levels Loss (ADLoss). ADLoss can adaptively divide the density map into three-level subgraphs.

We implement ADLoss as follows: (1) We divide the ground-truth into three-level subregions including low-density, medium-density and high-density, and denote the subregion by Si with $i \in \{1, 2, 3\}$. When the crowd count of the subregion Si is lower than a given threshold T1 or higher than threshold T2, the region is divided into low-density or high-density subregion and is labeled S1 or S3. And the other region is the medium-density subregion labeling S2. (2) We consider the relative estimated loss of each subregion and sum them to reach the $L_{ADLoss}$:

$$L_{ADLoss} = k1 * S1_{loss} + k2 * S2_{loss} + k3 * S3_{loss}. \tag{6}$$

There k1, k2, k3 represent the loss coefficient of different subregions, respectively. Note that: $\sum ki = 1$(i = 1, 2, 3).

**Table 1**
Comparison of architectures of the DDB module and the context-guided module on ShanghaiTech Part_B dataset.

| Architecture | SH Tech Part_B | |
|---|---|---|
| | MAE | MSE |
| Vgg16+1DDB | 10.5 | 16.1 |
| Vgg16+2DDB | 8.8 | 15.3 |
| Vgg16+3DDB | 8.3 | 14.5 |
| Vgg16+4DDB | 8.1 | 14.1 |
| Vgg16+Context+1DDB | 9.2 | 15.7 |
| Vgg16+Context+2DDB | 8.5 | 14.9 |
| Vgg16+Context+3DDB | 8.1 | 14.2 |
| Vgg16+Context+4DDB | **8.0** | **13.8** |

**Table 2**
Comparison of architectures of the DADBs and the context-guided module on ShanghaiTech Part_B dataset.

| Architecture | SH Tech Part_B | |
|---|---|---|
| | MAE | MSE |
| Vgg16+1DADB | 7.7 | 13.1 |
| Vgg16+2DADB | 7.3 | 12.7 |
| Vgg16+3DADB | 6.8 | 10.9 |
| Vgg16+4DADB | 6.6 | 10.5 |
| Vgg16+Context+1DADB | 7.2 | 11.6 |
| Vgg16+Context+2DADB | 6.9 | 11.1 |
| Vgg16+Context+3DADB | 6.6 | 10.5 |
| Vgg16+Context+4DADB | **6.5** | **10.2** |

## 4. Experiments

We evaluate our model on the ShanghaiTech(Part_A and Part_B) dataset [1], the UCF_CC_50 datasets [25], the UCSD dataset [26], and the WorldExpo'10 dataset [27]. To demonstrate the effectiveness of every module in CDADNet, we implement ablation experiments on the ShanghaiTech Part_B dataset. Then, we compare the proposed method with other existing approaches on these five datasets, and the following two standard evaluation metrics are used: MAE (Mean Absolute Error), MSE (Mean Squared Error) [3,15]. The MAE and MSE are defined as follows:

$$MAE = \frac{1}{\nu}\sum_{i=1}^{\nu}|\zeta_i - \hat{\zeta}_i| \tag{7}$$

$$MSE = \sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}\|\zeta_i - \hat{\zeta}_i\|^2} \tag{8}$$

where $\nu$ represents the volume of the test dataset images, $\zeta_i$ and $\hat{\zeta}_i$ are the $i$th ground-truth and generated density map, respectively.

### 4.1. Training details

We utilize a plain way to train the CDADNet. There we use the first 13 layers that are from a pre-trained VGG-16 to generate a fundamental density map [28], and the output from CDADNet is the generated density map. In other layers, the original values are derived from a Gaussian initialization with a mean zero and a standard deviation(std) 0.01. Adam optimizer with a learning rate of 1e-6 is leveraged to train our model. Furthermore, we select the Euclidean distance to estimate the error between the ground truth and the generated density map. The fulfillment of our training and test for the proposed architecture is based on the Pytorch framework.

### 4.2. Ablation study

In this section, we first present the generation of the attention map in the attention module. Fig. 4 shows the generation of the attention map on all these five datasets. Then, we performed an ablation study to analyze the configurations of the proposed CDADNet on the ShanghaiTech Part_B dataset. For a fair comparison, a Vgg-16 is used to build a front-end network. As shown in Table 1, the Dense Dilated Block (DDB) module and the Context-guide module significantly improve performance. The DDB module and the context module achieve improvement when the number of DDB module increase from 1 to 4. However, their effect may not be so obvious once the network deepens. Moreover, Table 2 presents that the Dense Attentional Dilated Blocks (DADBs) obtain better performance than DDB. The network containing 4 DADBs reduces the MAE of ShanghaiTech Part_B from 8.0 to 6.5 compared with containing 4 DDBs. Diverse performances in the same density map can be achieved by using different numbers of DADBs, as shown in Fig. 5.

**Table 3**
Loss on ShanghaiTech Part_B dataset.

| Loss | | MAE | MSE |
|---|---|---|---|
| Standard Loss | | 6.8 | 14.6 |
| AD Loss | Two Level | 6.6 | 12.4 |
| | Three Level | 6.5 | 10.2 |
| | Four Level | **6.5** | **9.8** |

**Table 4**
Estimation errors on the ShanghaiTech Part_A and Part_B datasets.

| Method | Part_A | | Part_B | |
|---|---|---|---|---|
| | MAE | MSE | MAE | MSE |
| Switching-CNN [2] | 90.4 | 135.0 | 21.6 | 33.4 |
| CP-CNN [15] | 73.6 | 106.4 | 20.1 | 30.1 |
| TDF-CNN [29] | 97.5 | 145.1 | 20.7 | 32.8 |
| Decidenet [30] | – | – | 20.8 | 29.4 |
| D-ConvNet-v1 [10] | 73.5 | 112.3 | 18.7 | 26.0 |
| CSRNet [3] | 68.2 | 115.0 | 10.6 | 16.0 |
| SANet [17] | 67.0 | 104.5 | 8.4 | 13.6 |
| TEDnet [31] | 64.2 | 109.1 | 8.2 | 12.8 |
| A+RR+SP [32] | 63.1 | 96.2 | 8.7 | 13.6 |
| ADCrowdNet(AbD) [33] | 63.2 | 98.9 | 8.2 | 15.7 |
| AT-CSRNet [12] | - | - | 8.1 | 13.5 |
| CAN [8] | 62.3 | 100.0 | 7.8 | 12.2 |
| PACNN+ [9] | 62.4 | 102.0 | 7.6 | 11.8 |
| DSNet [4] | 61.7 | 102.6 | 6.7 | 10.5 |
| ASD [34] | 65.6 | 98.0 | 8.5 | 13.7 |
| CAN [8] | 62.3 | 100.0 | 7.8 | 12.2 |
| PGCNet [6] | 57.0 | 86.0 | 8.8 | 13.7 |
| HA-CCN [35] | 62.9 | 94.9 | 8.1 | 13.4 |
| RANet [5] | 59.4 | 102.0 | 7.9 | 12.9 |
| DANet+ASNet [36] | 57.78 | 90.13 | – | – |
| CDADNet | **57.3** | **89.8** | **6.5** | **10.2** |

Moreover, in order to validate the validity of ADLoss, we test two-level, three-level, and four-level ADLoss, respectively. The experimental results are exhibited in Table 3 and show that the ADLoss further improves the counting performance of CDADNet. CDADNet with three-level ADLoss achieves an MAE of 6.3, outperforming standard loss and the net with two-level ADLoss. In addition, four-level ADLoss has a little improvement over three-level, but it costs a lot of computing resources. Generally speaking, the three-level ADLoss has a better performance and effect. Besides, a statistical analysis makes clear that our approach decreases the predicted loss in multiple density level regions. The results are shown in Fig. 6.

Compared with the counting from ground truth, different density level counting has different error values. To further reduce the errors of high-density, we compute $L_{ADLoss}$ by setting the values of k1, k2, and k3 to 0.3, 0.3, and 0.4, respectively.

### 4.3. Evaluation and comparison

#### 4.3.1. ShanghaiTech dataset

This dataset contains 1198 annotated images with a total amount of 330,165 persons [27]. It consists of two parts: Part_A contains 482

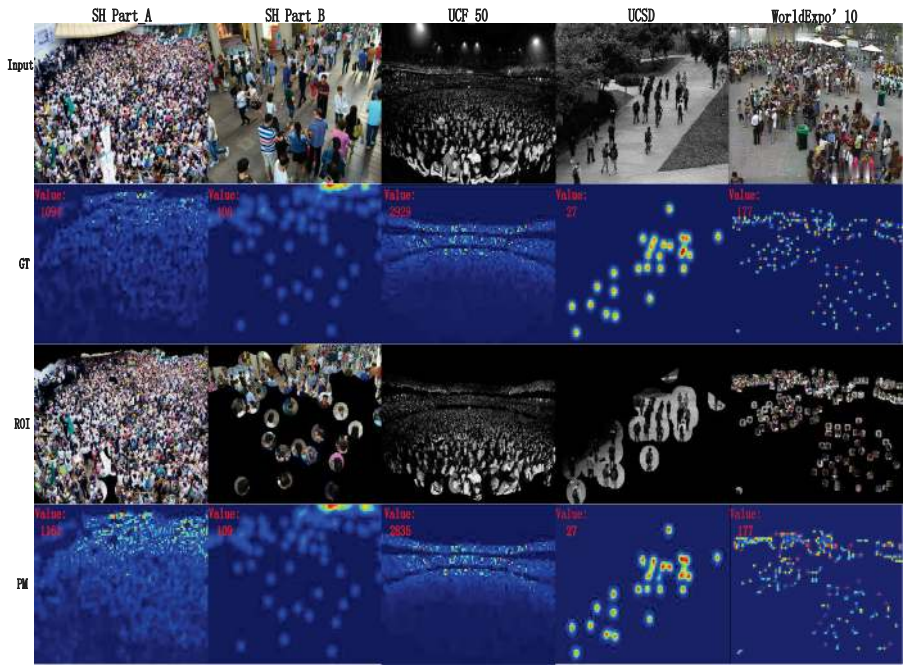



**Fig. 4. The results of Dense Attentional Dilated Block Network.** The attention map on five datasets. The first row shows five sample images from ShanghaiTech Part_A, ShanghaiTech Part_B, The WorldExpo'10, The UCSD, The UCF_CC_50 datasets. The second row shows ground truths (GTs) originating from datasets. The next row shows the regions of interest (ROI) multiplying input images and corresponding attention maps. The last row shows predicted maps (PMs) outputting from CDADNet.

**Table 5**
Estimation errors on the WorldExpo'10 dataset.

| Method | The WorldExpo'10 | | | | | |
|---|---|---|---|---|---|---|
| | Sce.1 | Sce.2 | Sce.3 | Sce.4 | Sce.5 | Ave. |
| Switching-CNN [2] | 4.4 | 15.7 | 10.0 | 11.0 | 5.9 | 9.4 |
| CP-CNN [15] | 2.9 | 14.7 | 10.5 | 10.4 | 5.8 | 8.9 |
| TDF-CNN [29] | 2.7 | 23.4 | 10.7 | 17.6 | 3.3 | 11.5 |
| Decidenet [30] | 2.0 | 13.1 | 8.9 | 17.4 | 4.8 | 9.2 |
| D-ConvNet-v1 [10] | 1.9 | 12.1 | 20.7 | 8.3 | 2.6 | 9.1 |
| CSRNet [3] | 2.9 | 11.5 | 8.6 | 16.6 | 3.4 | 8.6 |
| SANet [17] | 2.6 | 13.2 | 9.0 | 13.3 | 3.0 | 8.2 |
| PGCNet [6] | 2.5 | 12.7 | 8.4 | 13.7 | 3.2 | 8.1 |
| TEDnet [31] | 2.3 | 10.1 | 11.3 | 13.8 | 2.6 | 8.0 |
| A+RR+SP [32] | 2.9 | 15.0 | 7.2 | 14.7 | 2.6 | 8.5 |
| PACNN [9] | 2.3 | 12.5 | 9.1 | 11.2 | 3.8 | 7.8 |
| AT-CSRNet [12] | 1.8 | 13.7 | 9.2 | 10.4 | 3.7 | 7.8 |
| ADCrowdNet [33] | 1.7 | 14.4 | 11.5 | 7.9 | 3.0 | 7.7 |
| CAN [8] | 2.9 | 12.0 | 10.0 | 7.9 | 4.3 | 7.4 |
| SANet+SPANet [37] | 2.3 | 12.3 | 7.9 | 12.9 | 3.2 | 7.7 |
| DANet+ASNet [36] | 2.22 | 10.11 | 8.89 | 7.14 | 4.84 | 6.64 |
| CDADNet | **1.5** | **8.7** | **6.3** | **6.8** | **1.8** | **5.0** |

**Table 6**
Estimation errors on the UCF_CC_50 dataset.

| Method | UCF_CC_50 | |
|---|---|---|
| | MAE | MSE |
| Switching-CNN [2] | 318.1 | 439.2 |
| CP-CNN [15] | 295.8 | 320.9 |
| TDF-CNN [29] | 354.7 | 491.4 |
| D-ConvNet-v1 [10] | 288.4 | 404.7 |
| CSRNet [3] | 266.1 | 397.5 |
| SANet [17] | 258.4 | 334.9 |
| TEDnet [31] | 249.4 | 354.5 |
| PACNN+ [9] | 241.7 | 320.7 |
| ADCrowdNet(AbD) [33] | 266.4 | 358.0 |
| PGCNet [6] | 244.6 | 361.2 |
| RANet [5] | 239.8 | 319.4 |
| SANet+SPANet [37] | 232.6 | 311.7 |
| CAN [8] | 212.2 | 243.7 |
| ASD [34] | 196.2 | 270.9 |
| DSNet [4] | 183.3 | 240.6 |
| GCINet [38] | 179.8 | 262.4 |
| DANet+ASNet [36] | 174.84 | 251.63 |
| CDADNet | **170.5** | **228.7** |

images of the high-density population, and Part_B contains 716 images of the low-density population in Shanghai streets. We have compared our method to other twenty recent approaches and the results are shown in Table 4. Data makes clear our proposed method achieves the lowest MAE (the highest accuracy) and MSE on part_B, and goodish MAE and MSE on part_A.

#### *4.3.2. WorldExpo'10 dataset*

The dataset includes 3980 annotated images derive from 1132 video succession taken by 108 monitors [27]. These labeled images take from five different scenes and our proposed architecture delivers the best accuracy in all scenes. Results are shown in Table 5. The proposed CDADNet significantly outperforms other excellent models on all of the five scenes.

#### *4.3.3. UCF_CC_50 dataset*

The dataset includes 50 images with various perspective and resolution [25]. The number of people per image ranges from 94 to 4543 and each picture contains an average of 1280 people. The comparative result of MAE and MSE are included in Table 6. It indicates that our method can obtain the lowest MAE and MSE. Hence, our method has a good performance for dense scenes.

#### *4.3.4. UCSD dataset*

The dataset [26] consists of 2000 annotated images captured from the sparse scenes. The number of annotated persons per image varies from 11 to 46. The training set contains images that index from 600 to 1399 and the test set contains remained 1200 images [38]. The accuracy of the UCSD dataset is shown in Table 7. Compared with excellent methods, the proposed CDADNet achieves the highest accuracy on both MAE and MSE.

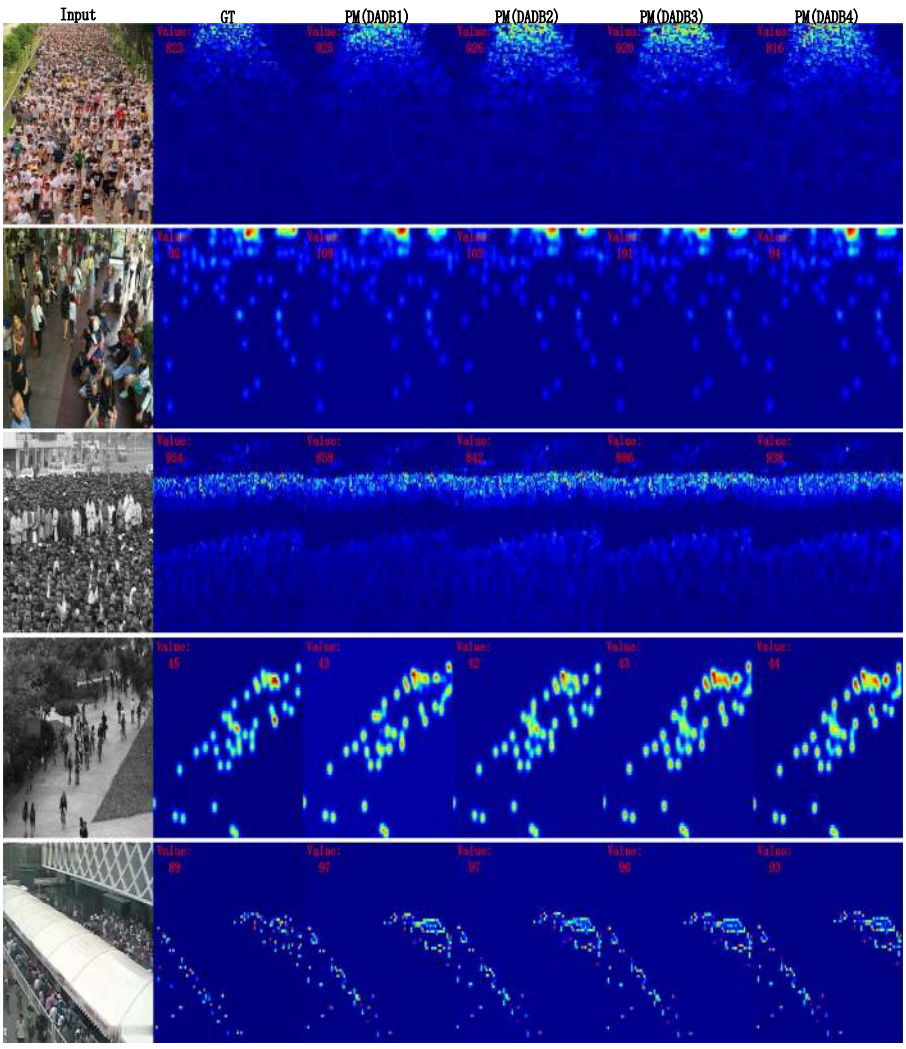



**Fig. 5. The density maps from CDADNet using 1,2,3,4 Dense Attentional Dilated Block.** The input images sample from the above-mentioned five datasets. From the first row to the last row are drawn from ShanghaiTech Part_A, ShanghaiTech Part_B, The UCF_CC_50, The UCSD , The WorldExpo'10 datasets. And from the first column to the last column are input images, corresponding ground truths, and output predicted maps (PMs) using 1–4 DADBs.

**Table 7**
Estimation errors on the UCSD dataset.

| Method | UCSD | |
|---|---|---|
| | MAE | MSE |
| Cross-Scene [27] | 1.60 | 3.31 |
| Switching-CNN [2] | 1.62 | 2.10 |
| CSRNet [3] | 1.16 | 1.47 |
| ACSCP [39] | 1.04 | 1.35 |
| SPN [40] | 1.03 | 1.32 |
| ADCrowdNet(AaD) [33] | 1.09 | 1.35 |
| SANet+SPANet [37] | 1.00 | 1.28 |
| SANet [17] | 1.02 | 1.29 |
| PACNN [9] | 0.89 | 1.18 |
| DSNet [4] | 0.82 | 1.06 |
| GCINet [38] | 1.14 | 1.43 |
| CDADNet | **0.75** | **1.02** |

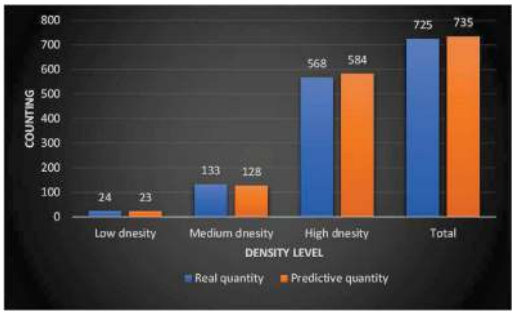



**Fig. 6. The counting of different levels density maps**. The network has a few errors in the low-density region and medium-density region. However, there is a big error in the high-density region.

## 5. Conclusion

In this paper, we design a Context-guided Dense Attentional Dilated Network (CDADNet) to generate high-granularity density maps. CDADNet contains three components: an attentional module, a context-guided module, and a Dense Attentional Dilated Module. The attentional module is used to provide attention maps, while the context-guided module is proposed to extract multi-scale contextual information. Moreover, the dense attentional dilated module is used to address the large-scale variations and complex background problems. We evaluate our approach on five popular datasets and the results indicate that our approach has achieved a brilliant performance. Thus, we could conclude that the proposed method has super excellent capacity on the

crowd counting tasks for extremely dense scenes and relatively sparse scenes.

### CRediT authorship contribution statement

**Aichun Zhu:** Conceptualization, Methodology, Writing- reviewing & editing. **Guoxiu Duan:** Software, Writing- original draft preparation. **Xiaomei Zhu:** Software, Visualization, Data curation. **Lu Zhao:** Writing- reviewing & editing. **Yaoying Huang:** Investigation, Validation. **Gang Hua:** Supervision. **Hichem Snoussi:** Supervision.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61972016 and 61802176), China Postdoctoral Science Foundation (Grant No. 2019M661999) and Natural Science Foundation of Jiangsu Higher Education Institutions of China (19KJB520009).

### References

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, 2016, pp. 589–597, http://dx.doi.org/10.1109/CVPR.2016.70.

[2] D.B. Sam, S. Surya, R.V. Babu, Switching convolutional neural network for crowd counting, 2017, pp. 4031–4039, http://dx.doi.org/10.1109/CVPR.2017.429.

[3] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, 2018, pp. 1091–1100, http://dx.doi.org/10.1109/CVPR.2018.00120.

[4] F. Dai, H. Liu, Y. Ma, J. Cao, Q. Zhao, Y. Zhang, Dense scale network for crowd counting, 2019.

[5] A. Zhang, J. Shen, Z. Xiao, F. Zhu, X. Zhen, X. Cao, L. Shao, Relational attention network for crowd counting, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2020.

[6] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, E. Ding, Perspective-guided convolution networks for crowd counting, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019.

[7] D. Ooro-Rubio, R.J. López-Sastre, Towards perspective-free object counting with deep learning, in: European Conference on Computer Vision, ECCV, 2016, pp. 615–629.

[8] W. Liu, M. Salzmann, P. Fua, Context-aware crowd counting, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 5094–5103., http://dx.doi.org/10.1109/CVPR.2019.00524.

[9] M. Shi, Z. Yang, C. Xu, Q. Chen, Revisiting perspective information for efficient crowd counting, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[10] Z. Shi, Z. Le, L. Yun, X. Cao, G. Zheng, Crowd counting with deep negative correlation learning, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[11] J. Wan, A. Chan, Adaptive density map generation for crowd counting, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 1130–1139, http://dx.doi.org/10.1109/ICCV.2019.00122.

[12] M. Zhao, J. Zhang, C. Zhang, W. Zhang, Leveraging heterogeneous auxiliary tasks to assist crowd counting, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 12736–12745, http://dx.doi.org/10.1109/CVPR.2019.01302.

[13] H. Zhang, I. Goodfellow, D.N. Metaxas, A. Odena, Self-attention generative adversarial networks, 2018, arXiv: Machine Learning.

[14] R.R. Varior, B. Shuai, J. Tighe, D. Modolo, Scale-aware attention network for crowd counting, 2019, CoRR abs/1901.06026. arXiv:1901.06026. URL http://arxiv.org/abs/1901.06026.

[15] V.A. Sindagi, V.M. Patel, Generating high-quality crowd density maps using contextual pyramid cnns, 2017, pp. 1879–1888, http://dx.doi.org/10.1109/ICCV.2017.206.

[16] L. Zhang, M. Shi, Q. Chen, Crowd counting via scale-adaptive convolutional neural network, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, 2018, pp. 1113–1121, http://dx.doi.org/10.1109/WACV.2018.00127.

[17] X. Cao, Z. Wang, Y. Zhao, F. Su, Scale aggregation network for accurate and efficient crowd counting, in: V. Ferrari, C. Sminchisescu M. Hebert and, Y. Weiss (Eds.), Computer Vision, ECCV 2018, Springer International Publishing, Cham, 2018, pp. 757–773.

[18] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692, http://dx.doi.org/10.1109/CVPR.2018.00388.

[19] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 842–850, http://dx.doi.org/10.1109/CVPR.2015.7298685.

[20] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A. Yuille, Attention to scale: Scale-aware semantic image segmentation, 2016, pp. 3640–3649, http://dx.doi.org/10.1109/CVPR.2016.396.

[21] W. Wang, S. Zhao, J. Shen, S. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, http://dx.doi.org/10.1109/CVPR.2019.00154.

[22] A. Zhu, Z. Zheng, Y. Huang, T. Wang, J. Jin, F. Hu, G. Hua, H. Snoussi, Cacrowdgan: Cascaded attentional generative adversarial network for crowd counting, IEEE Trans. Intell. Transp. Syst. (2021) 1–13, http://dx.doi.org/10.1109/TITS.2021.3075859.

[23] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3141–3149, http://dx.doi.org/10.1109/CVPR.2019.00326.

[24] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6450–6458, http://dx.doi.org/10.1109/CVPR.2017.683.

[25] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, 2013, pp. 2547–2554, http://dx.doi.org/10.1109/CVPR.2013.329.

[26] A. Chan, Z.-S. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[27] Z. Cong, H. Li, X. Wang, X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv:1409.1556.

[29] D. Sam, R. Babu, Top-down feedback for crowd counting convolutional neural network, 2018.

[30] J. Liu, C. Gao, D. Meng, A. Hauptmann, Decidenet: Counting varying density crowds through attention guided detection and density estimation, 2018, pp. 5197–5206, http://dx.doi.org/10.1109/CVPR.2018.00545.

[31] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, L. Shao, Crowd counting and density estimation by trellis encoder–decoder networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 6126–6135, http://dx.doi.org/10.1109/CVPR.2019.00629.

[32] J. Wan, W. Luo, B. Wu, A.B. Chan, W. Liu, Residual regression with semantic prior for crowd counting, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 4031–4040, http://dx.doi.org/10.1109/CVPR.2019.00416.

[33] N. Liu, Y. Long, C. Zou, Q. Niu, H. Wu, Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[34] X. Wu, Y. Zheng, H. Ye, W. Hu, L. He, Adaptive scenario discovery for crowd counting, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019.

[35] V.A. Sindagi, V.M. Patel, Ha-ccn: Hierarchical attention-based crowd counting network, IEEE Trans. Image Process. 29 (2020) 323–335, http://dx.doi.org/10.1109/TIP.2019.2928634.

[36] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, Y. Pang, Attention scaling for crowd counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[37] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, A. Hauptmann, Learning spatial awareness to improve crowd counting, 2019.

[38] Y. Wang, X. Wang, X. Bai, T. Zhao, Y. Cheng, Global context instructive network for extreme crowd counting, IEEE Access 8 (2020) 34265–34275, http://dx.doi.org/10.1109/ACCESS.2019.2962870.

[39] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5245–5254, http://dx.doi.org/10.1109/CVPR.2018.00550.

[40] X. Chen, Y. Bin, N. Sang, C. Gao, Scale pyramid network for crowd counting, 2019, pp. 1941–1950, http://dx.doi.org/10.1109/WACV.2019.00211.