Big Data
Volume 00, Number 00, 2022

DOI: 10.1089/big.2022.0039

Spatial-Frequency Attention Network for Crowd Counting

Xiangyu Guo,[1] Mingliang Gao,[1,*] Wenzhe Zhai,[1] Jianrun Shang,[1] and Qilei Li[2]

Abstract

Counting the number of people in crowded scenarios is a crucial task in video surveillance and urban security system. Widely deployed surveillance cameras provide big data for training, a compelling deep learning-based counting network. However, large-scale variations in dense crowds are still not entirely solved. To address this problem, we propose a spatial-frequency attention network (SFANet) for crowd counting in this article. A bottleneck spatial attention module is built to emphasize features in various spatial locations and select a region containing individuals adaptively in the spatial domain. As a complementary, in the frequency domain, a multispectral channel attention module is adopted to obtain a more complete set of frequency components for representing each channel. The two attention modules are combined to focus on the discriminative region and suppress the misleading information by their mutual promotion. Experimental results on five benchmark crowd data sets demonstrate that the SFANet can achieve the state-of-the-art performance in terms of accuracy and robustness.

Keywords: crowd counting; density estimation; spatial-frequency attention; convolutional neural network

Introduction

Crowd counting aims to estimate the total number of people presented in a crowded scene. Owing to its wide range of real-world applications, for example, traffic control[1] and public safety,[2] crowd counting has drawn much attention. The widely deployed surveillance cameras provides large-scale data to train a compelling crowd counter, which significantly boost the performance in recent years. However, various challenges, such as scale variations, perspective distortion, serious occlusion, and nonuniform distribution, hinder the further performance improvement of current methods.

To tackle the aforementioned problems, various methods have been proposed in recent years. Early works utilize detection and regression methods for crowd counting.[3,4] However, these two methods perform unsatisfactorily in congested scenes.[5,6] With the raising of deep learning, an increasing number of Convolution Neural Networks (CNNs)-based counting methods have been proposed.[7,8] These approaches conducted crowd counting by learning density maps in an end-to-end manner. Meanwhile, inspired by the success of attention mechanism in visual tasks,[9,10] the attention mechanism has been widely explored for crowd counting.[11–13] Although the aforementioned methods have been proposed for crowd counting, they are still helpless for scale variations. The scale variations are caused by camera perspective distortion, which results in different distances between heads and cameras. Figure 1 depicts some congested scenes in which the crowds are suffered from scale variations.

In this article, we propose a spatial-frequency attention network (SFANet) to solve the large-scale variations in dense crowds scenes. The proposed method consists of two attention units, which separately handle information in spatial and frequency domains, whereas their combinations proved each other with the complementary information to exploit the inherent synergy to felicitate a precious crowd counting. Specifically, in the spatial domain, a bottleneck spatial attention (BSA) module is built to emphasize features in various spatial

[1]School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom.

*Address correspondence to: Mingliang Gao, School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, E-mail: mlgao@sdut.edu.cn

FIG. 1. The scale variations in congested scenes. The green box denotes the head region.

locations and select a region containing individuals adaptively. Thus, it alleviates the mistaken estimation for background regions. In the frequency domain, a multispectral channel attention (MSCA) module is designed to consider a more complete set of frequency components for representing each channel. In this case, the people rather in the crowd can be recognized accurately. The contributions of the proposed method can be generally summarized as follows:

1. We propose a crowd counting network with two elaborately designed attention modules, BSA module and MSCA module, to capture more information in spatial and channel dimensions.
2. We analyze crowd counting in the spatial and frequency domain simultaneous to exploit the inherent synergy. In spatial domain, a larger receptive field is utilized to determine the person region. Meanwhile, multiple frequency components are used to identify people in the frequency domain.
3. We perform comprehensive experiments on five crowd benchmark data sets with other state-of-the-art methods to demonstrate the superiority of the proposed method in terms of both accuracy and robustness.

Related Work

Progressive improvements in crowd counting have been made from traditional methods[3,5] to CNN-based method.[14,15] In this section, we mainly retrospect these two methods.

Traditional approaches

Traditional approaches can be mainly divided into two categories: detection-based and regression-based approaches.[7,8] The detection-based method mainly employs a sliding-window-like detector to detect the body or head of each individual, and then trains a classifier to determine the positive instances.[16] Li et al.[17] employed a Histogram of Oriented Gradient-based head-shoulder detector to count people in foreground regions. Ge and Collins[18] used a Bayesian marked point process to detect people. However, the method performed unsatisfactory in dense crowds. The detection-based approaches usually perform well in sparse scenes, but the performance degrades seriously in the high-density crowd scenarios.[7]

To deal with the problem of counting in high-density crowd scenarios, the regression-based approaches have been proposed. The regression-based approaches mainly learn a mapping from an image to the count. They first extract global or local features, and leverage a regressor to learn a mapping for crowd counting. Lempitsky and Zisserman[19] regressed a density map to leverage spatial distribution information, which learned a linear mapping between local features and density maps. Chan and Vasconcelos[20] introduced a prior distribution on the linear function to explain the Poisson regression based on Bayesian. Although the two traditional approaches achieve promising results, they always ignore spatial information.

CNN-based approaches

Recently, benefiting from the powerful ability of feature expression and computing resources, the CNN-based methods[7] have achieved great success in crowd counting. Zhang et al.[21] presented multicolumn CNN (MCNN) to solve the scale variations by increasing the receptive field. Similarly, a switch-CNN[22] employed multicolumn structure to train independent CNN regressors and selected the best classifier for estimation. However, the multicolumn structures usually lead to much information redundancy and consume a lot of parameters. To address these problems, Li et al.[15] employed a single-column network architecture with cascaded dilated convolutional layers to extract multi-scale features. More recently, many advanced approaches have exploited multi-context information to deal with the scale variation problem. Cheng et al.[23]

proposed a decoupled two-stage counting network that sequentially regresses the probability map and counter map.

Inspired by the human visual attention, attention mechanisms have gained comprehensive attentions in the field of deep learning.[24] Attention mechanism can be regarded as a process of dynamic weight adjustment. Some works have incorporated attention mechanisms in crowd counting to enhance the performance in recent years. Liu et al.[25] combined the visual attention mechanism and multiscale deformable convolutional scheme into a cascading framework to provide regions and congestion degrees for the latter density map estimator. Sindagi and Patel[26] employed the spatial attention module and global attention module to enhance the features adaptively. Jiang et al.[27] proposed an attention scaling network to alleviate the counting performance differences in different regions. Unlike the aforementioned attention-based methods, which only use the attention mechanism in the spatial domain, we build a hybrid-attention model to capture more information in the spatial domain and in the frequency domain.

The Proposed Approach

Overview

Considering the scale variations in extremely dense crowds, it is crucial to determine the discriminative area of a person, regardless of the view point and perspective distortion. To this end, we propose the SFANet to highlight the discriminative information for crowd counting by adaptively adjusting the weights despite the scale variation. To precisely locate the region of people to alleviate the mis-estimation caused by background region, we first propose the BSA module. Moreover, as pointed out in Qin et al.,[9] only the lowest frequency information is preserved and the high-frequency components are discarded in the traditional channel attentions. To overcome this inherent drawback, we propose the MSCA module to take the advantage of the high-frequency components to represent the details of a person. The architecture of the proposed method is illustrated in Figure 2.

The proposed SFANet takes a pretrained ResNet-50 as feature extractor. Then, the feature maps are fed into BSA units to generate a two-dimensional spatial map $M_s \in R^{1\times H\times W}$, which adaptively adjusts weight in spatial dimension. It can be represented by

$$F_s = O_s(F) \otimes F, \tag{1}$$

where F_s is the enhanced feature map. $O_s(\cdot)$ denotes the function of BSA module and $\otimes$ is the element-wise multiplication, and F is the output of feature extractor. Furthermore, the feature map F_s is refined by MSCA module to take the advantage of the information in the frequency domains $V_c \in R^{C\times 1\times 1}$ in channel dimension. The refined feature map can be written as

$$F_c = O_c(F_s) \otimes F_s, \tag{2}$$

where F_c is the optimized feature map and $O_c(\cdot)$ denotes the function of MSCA module. Finally, the feature map $F_{cs} \in R^{C\times H\times W}$ is obtained by the sum F_s and F_c.

$$F_{cs} = F_s \oplus F_c, \tag{3}$$

where $\oplus$ is a sum operation. In this case, F_{cs} can accurately determine the discriminative area of a person.

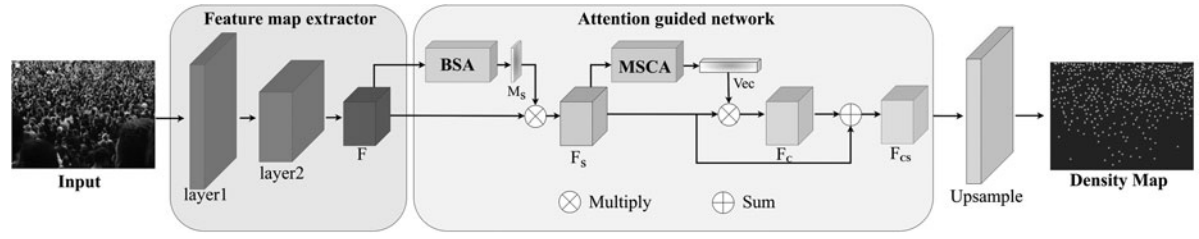

FIG. 2. The architecture of the proposed SFANet. A feature map extractor is designed to extract feature. Two attention modules, BSA and MSCA, are built to adjust weights in spatial and channel dimension, respectively. These two types of feature maps are concatenated and produces a one-channel predicted density map through upsample operation. BSA, bottleneck spatial attention; MSCA, multispectral channel attention.

BSA module

Spatial attention devotes to highlight or restrain features in various spatial locations by producing a spatial attention map. To select the head regions, more contextual information needs to be leveraged, thus a larger receptive field is required. To enlarge the receptive field, a BSA unit[28] is built in this article. In BSA unit, the dilated convolution is adopted to construct a representative feature map, which ensures that there is a large enough receptive field to select a discriminative area. Beyond that, the BSA is designed as a bottleneck structure to reduce the computational costs. The architecture of the BSA unit is shown in Figure 3.

First, to enable the feature map to be more compact, we apply a 1×1 convolution upon the extracted feature map F to reduce the channel dimension to C/r, where r represents the reduction ratio (r is set to 16 in this study). Subsequently, two 3×3 dilated convolutions are employed to take more contextual information into account. Finally, the channel of the feature map is further refined to one dimension with a 1×1 convolution, and a spatial attention map $M_s \in \mathbf{R}^{1\times H\times W}$ is obtained. The BSA module can be formulated as

$$M_s(X) = \mathrm{BN}(\mathrm{Conv}_2^{1\times1}((\mathrm{DC}_{1,2}^{3\times3}(\mathrm{Conv}_1^{1\times1}(X))))), \quad (4)$$

where BN represents the batch normalization operation. $\mathrm{Conv}_2^{1\times1}$ and $\mathrm{Conv}_2^{1\times1}$ are designed for channel reduction. $\mathrm{DC}_{1,2}^{3\times3}$ represents the two dilated convolution layers.

MSCA module

Channel attention aims to represent and evaluate the significance of each channel using a weight. From the perspective of the frequency domain, the dominated component of an image is low frequency. It forms the basic gray level of the image. Whereas the intermediate-frequency information forms the main edge structure of the image. In contrast, the high-frequency information forms the edges and details of the image.

To identify a person accurately in the discriminative region, all frequency components are expected to be exploited, not just the lowest one, as mentioned in Jiang et al.[27] To this end, we hereby employ the MSCA module, which leverages the discrete cosine transform (DCT) upon the traditional Global Average Pooling module, to sententiously consider information in all frequencies. The DCT is defined as

$$F(u,v) = \sum_{i=1}^{H}\sum_{j=1}^{W} f(i,j)\cos\left[\frac{(i+0.5)\pi}{H}h\right]\left[\frac{(j+0.5)\pi}{W}w\right], \quad (5)$$

where $F(i,j)$ represents the two-dimensional DCT frequency spectrum and $f(i,j)$ is the input. H and W denote the height and width of $f(i,j)$. The architecture of the MSCA is shown in Figure 4.

The MSCA module can be represented as

$$V^i = \mathrm{DCT}^i \otimes X^i,$$
$$M_c(F_s) = \mathrm{Sigmoid}(Fc(\mathrm{Cat}(V^i))),$$

where $V^i \in R^{C'\times1\times1}$, $i \in \{1,2,\cdots,n\}$, $C' = \frac{C}{n}$ denotes a channel vector. $\otimes$ denotes the element-wise multiplication. DCT^i is a selected frequency component (referred to DCT bases in this article) by the criteria[9] and X^i represents the part that is split along the

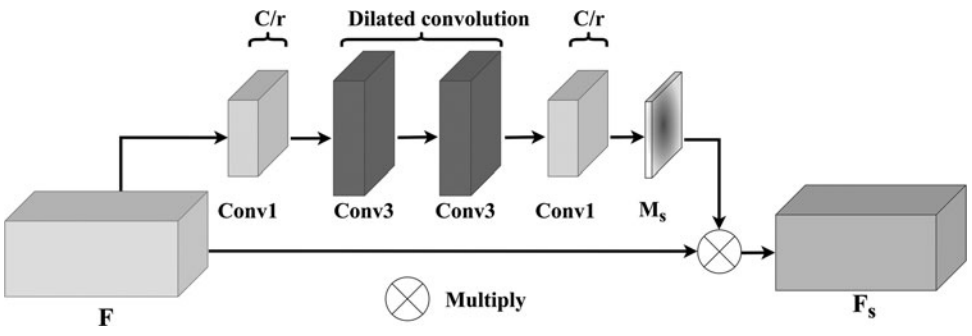

FIG. 3. Architecture of BSA module. It adopts a 1×1 convolution to reduce channels, and two dilated convolutions to enlarge the receptive fields. Finally, the spatial attention map is refined by a 1×1 convolutional layer.

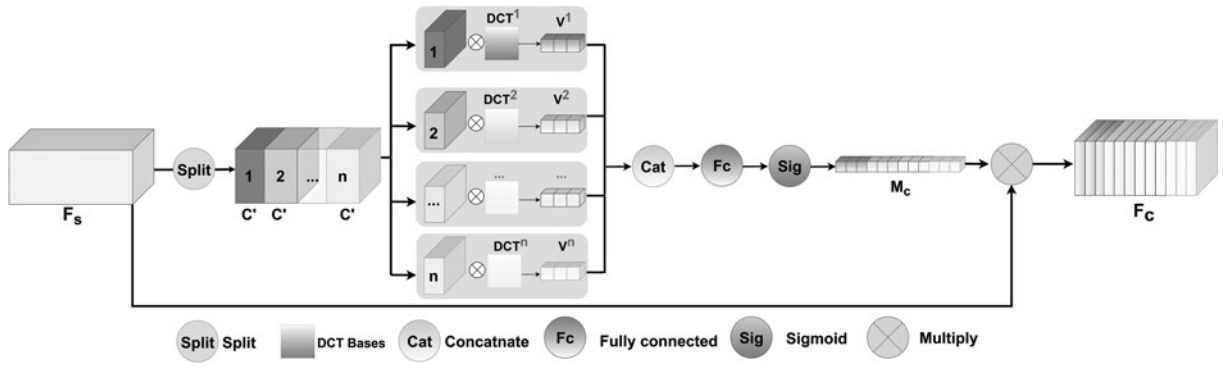

FIG. 4. Architecture of the MSCA module. The high-level feature F_s is split into n parts, on which the multiplication of each part and DCT bases is applied to obtain the optimized feature V^i. Finally, a channel representational vector is generated by the concatenation operation. DCT, discrete cosine transform.

channel dimension by the input F_s. $FC(\cdot)$ is the full connection operation. By this means, the proposed method can take the information in all the frequencies to produce a refined feature. Thus, the feature map F_c can be more accurate in identifying individuals.

Loss function

We use L_2 loss function to minimize the root-mean-square error (RMSE) between the ground truth (GT) and the predicted density map. Given an image I_i and the learnable parameter θ of SFANet, the goal is to minimize the following loss function:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}||\text{Est}(X_i,\theta) - \text{GT}_i||_2^2, \tag{7}$$

where N is the batch size of the input images in a mini-batch, $\text{Est}(X_i,\theta)$ denotes the estimated density map, and GT_i is the corresponding GT.

Density map generation

The GT density map $M(x)$ is generated by the geometry-adaptive Gaussian kernel G_σ[21] and convolving with a delta function. The formula is defined as follows:

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

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 employed to depict a head. It is equal to 1 when the pixel i is in a head region.

Experiments and Analysis

Implementation details

In this study, the training and evaluation are implemented on an NVIDIA RTX3090 GPU using PyTorch framework.[29] We employ Adam[30] as the optimizer. The learning rate is initialized as 10^{-5} and reduced to $\times 0.995$ per epoch to minimize the training loss.

Evaluation metrics

Following the previous works,[15,31] the mean absolute error (MAE) and RMSE are adopted to measure the performance of the proposed method. The MAE and RMSE are formulated in Equations (9) and (10), respectively.

Table 1. Experimental results on the ShanghaiTech data set

	Part_A		*Part_B*	
Method	*MAE*	*RMSE*	*MAE*	*RMSE*
GP[42]	120.4	179.4	12.5	18.3
MCNN[21]	110.2	173.2	26.4	41.3
CMTL[31]	101.3	152.4	20.0	31.1
TDF-CNN[43]	97.5	145.1	20.7	32.8
Switching-CNN[22]	90.4	135.0	21.1	30.1
CP-CNN[44]	73.6	106.4	20.1	30.1
BSAD[45]	90.4	135.0	20.2	35.6
TDF-CNN[43]	97.5	145.1	20.7	32.8
SaCNN[46]	86.8	139.2	20.7	32.8
A-CCNN[47]	85.4	124.6	11.0	19.0
MATT[48]	80.1	129.4	11.7	17.5
MRA-CNN[34]	74.2	112.5	11.9	21.3
PCC-Net[30]	73.5	124.0	19.2	31.5
DNCL[33]	73.5	**112.3**	18.7	26.0
SFANet(ours)	**71.7**	122.5	**8.6**	**13.7**

The best result is in bold.
MAE, mean absolute error; MCNN, multicolumn CNN; RMSE, root-mean-square error; SFANet, spatial-frequency attention network.

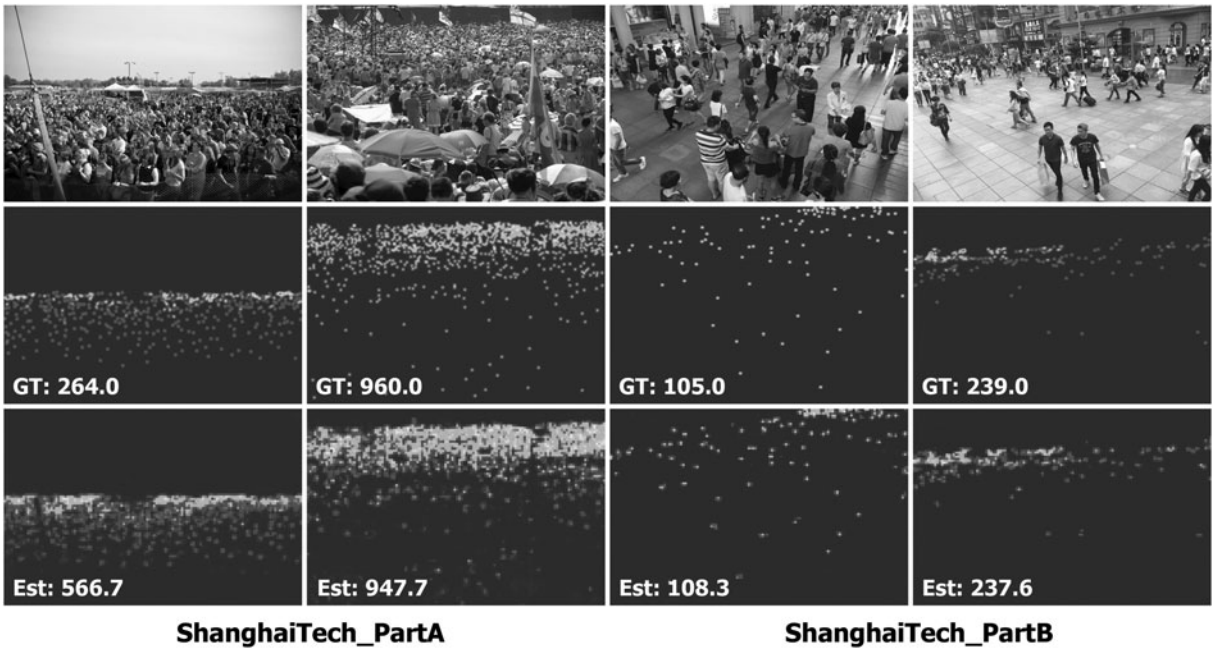

FIG. 5. Visually qualitative analysis of the ShanghaiTech data set. The top row shows the input images from Part_**(A)** and Part_**(B)**, the middle row shows the ground truth, and the bottom row shows the estimated density map.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_i^{\mathrm{pred}} - C_i^{gt}\right|, \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i^{\mathrm{pred}} - C_i^{gt}\right|^2}, \tag{10}$$

where N is the total number of testing images, C_i^{pred} is the predicted count value of the ith testing image, and C_i^{gt} is the corresponding GT count.

Data augmentation

To avoid the risk of overfitting and ensure the network being sufficiently trained, we augment the training data with random cropping and horizontal flipping instead of vertical flipping, because the vertical flipping reverse the positions of the head and feet, which is unfit for counting accuracy. Following the principle in C^3,[29] we constrain the input size to guarantee that they are divisible by 16. For ShanghaiTech, University of Central Florida-Qatar National Research Fund and University of Central Florida_Crowd Counting_50 data sets, we set the height and width of input images to the dimension of 768×1024. For WorldExpo'10 and NWPU-Crowd, we set the input images to 576×720 and 576×768, respectively. This constraint ensures that some max-pooling layers could output expected size.

Performance on ShanghaiTech data set.

The ShanghaiTech data set[21] is one of the most popular crowd counting data sets. It contains 1198 images with 330,165 annotated persons. This data set is divided

Table 2. Experimental results on the UCF-QNRF data set

Methods	*MAE*	*RMSE*
Zhang et al.[36]	467.0	498.5
Idress et al.[6]	315.0	508.0
MCNN[21]	277.0	509.1
CMTL[31]	252.0	514.0
Switching-CNN[22]	228.0	445.0
PCCNet[32]	148.7	247.3
CL[35]	132.0	191.0
CSRNet[15]	129.0	209.0
DENet[49]	121.0	205.0
LSC-CNN[50]	120.5	218.2
HA-CCN[26]	118.1	180.4
DUBNet[51]	116.0	**178.0**
DADNet[14]	113.2	189.4
SFANet(ours)	**111.3**	195.5

The best result is in bold.
UCF-QNRF, University of Central Florida-Qatar National Research Fund.

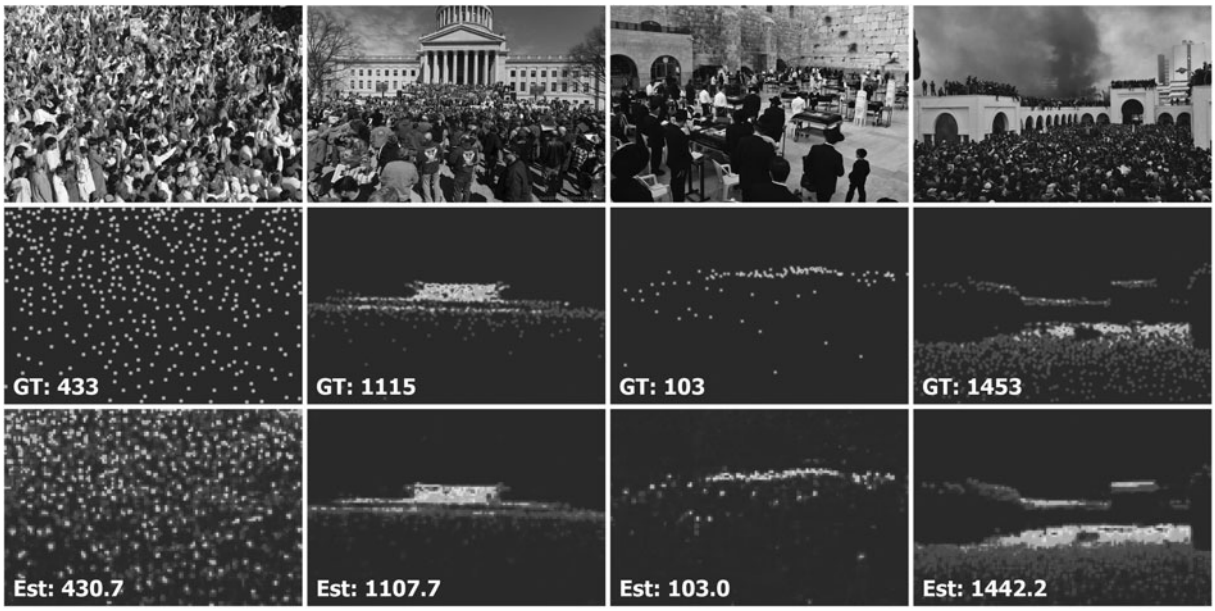

FIG. 6. Visually qualitative analysis of the UCF-QNRF data set. The first, second, and third row show the input images, ground truth density maps, and estimated density maps, respectively. UCF-QNRF, University of Central Florida-Qatar National Research Fund.

into two subsets, Part_A and Part_B. Part_A contains 482 images (300 images for training and 182 images for testing) randomly collected from internet. Part_B includes 716 images (400 images for training and 316 images for testing) captured from a congested street in Shanghai. By contrast, the images in Part_B have a smaller intra-data set divergence. The experimental results on the ShanghaiTech data set are shown in Table 1.

It can be seen that the proposed method scores 71.7 and 122.5 in terms of MAE and RMSE on Part_A sub-data set. Especially, it ranks the first place in MAE and improves the MAE by 2.5% compared with the second-best method, PCC-Net[32] and DNCL.[33] For Part_B, the proposed method scores 8.6 and 13.7 in MAE and RMSE, which both outperform the other competitors. Particularly, it reduces the MAE and RMSE by 27.7% and 35.7% compared with Multi-Resolution Attention-CNN,[34] which also leverages attention module. Some examples of visualization are shown in Figure 5. It shows that the proposed method performs well in the scenarios with scale variations.

Performance on UCF-QNRF data set. The UCF-QNRF data set[35] consists of 1535 high-resolution images with 1,251,642 individuals. The training set is composed of 1201 images and the testing set contains 334 images. Particularly, it has a wider variety of scenes compared with the ShanghaiTech data set. Comparative results are shown in Table 2. It shows that the SFANet scores 111.3 in MAE, which ranks the first place, and 195.5 in RMSE, which ranks the fifth place. The experimental results indicate that the proposed method is superior to other methods in MAE and remains competitive in

Table 3. Experimental results on the UCF_CC_50 data set

Methods	*MAE*	*RMSE*
MATT[48]	355.0	550.2
DR-ResNet[52]	307.4	421.6
CSRNet[15]	266.1	397.5
ic-CNN[53]	260.9	365.5
SCAR[40]	259.0	374.0
HA-CNN[26]	256.2	348.4
DM-Count[54]	211.0	291.5
AMSNet[55]	208.4	297.3
ADSCNet[41]	198.4	267.3
AMRNet[56]	184.0	265.8
TopoCount[57]	184.1	258.3
D2CNet[23]	182.1	254.9
LibraNet[58]	181.2	262.2
SFANet (ours)	**179.5**	**231.5**

The best result is in bold.
UCF_CC_50, University of Central Florida_Crowd Counting_50.

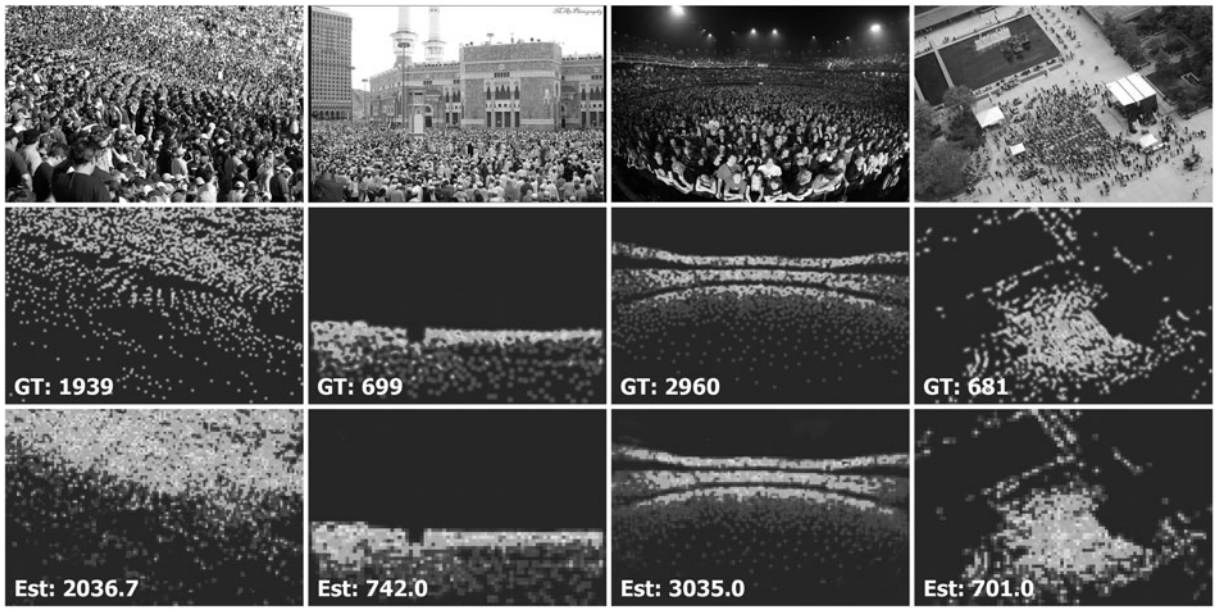

FIG. 7. Visually qualitative analysis of the UCF_CC_50 data set. The first, second, and third row show the input images, ground truth density maps, and estimated density maps, respectively. UCF_CC_50, University of Central Florida_Crowd Counting_50.

RMSE. The qualitative results for sample images from the UCF-QNRF data set is illustrated in Figure 6. It proves that the SFANet performs well under large-scale variations.

Performance on UCF_CC_50 data set. The UCF_CC_50 data set[6] contains 50 images with 63,947 annotated heads, which are collected from internet. The data set consists of packed scenes and the numbers of annotations range from 94 to 4543. The average annotations of one image reach 1280. It is a very challenging data set for the limited training samples and high-density crowds. The experimental results on the ShanghaiTech data set are shown in Table 3. It shows that the proposed SFANet scores 179.5 in MAE and 231.5 in RMSE. These two indicators surpass all the other competitors. Specifically, compared with the Hierarchical Attention-CNN,[26] which also adopts the attention mechanism in crowd counting, the proposed SFANet reduces the score of MAE by 29.9%, and RMSE by 33.6%, respectively. The visualization of the estimated crowd density maps with counting number is depicted in Figure 7. It proves that the estimated crowd density maps and counting values are approximate to the GT in high-density crowd scenario.

Performance on WorldExpo10 data set. The WorldExpo10 data set[36] is a large-scale crowd counting data set collected from Shanghai 2010 WorldExpo. It contains 3980 images, among which 3380 frame annotations are used for training, whereas 600 frames are used for testing. Since five different regions of interest (ROI) and the perspective maps are provided for the test scenes (S1–S5), we count persons within the ROI area following the general criterion.[37,38] The performance of the proposed SFANet against the state-of-

Table 4. Experimental results on the WorldExpo'10 data set

Methods	*Scene 1*	*Scene 2*	*Scene 3*	*Scene 4*	*Scene 5*	*MAE (Avg.)*
DCL[59]	1.8	16.2	9.2	25.0	2.8	11.0
C-CNN[60]	3.8	20.5	8.8	8.8	7.7	9.9
PCC Net[32]	1.9	18.3	10.5	13.4	3.4	9.5
DNCL[33]	1.9	12.1	20.7	8.3	2.6	9.1
ACM-CNN[61]	2.4	10.4	11.4	15.6	3.0	8.6
RPNet[62]	2.4	10.2	9.7	11.5	3.8	8.2
CRNet[63]	2.0	10.6	12.2	**7.8**	2.7	7.1
DENet[49]	2.8	10.7	8.6	15.2	3.5	8.2
SDANET[64]	2.0	14.3	12.5	9.5	2.5	8.1
ANF[65]	2.1	10.6	15.1	9.6	3.1	8.1
MRA-CNN[34]	2.4	11.4	9.3	10.5	3.7	7.5
CAT-CNN[66]	2.2	**9.8**	10.2	11.2	2.5	7.2
SFANet(ours)	**0.9**	12.0	**8.5**	10.0	**2.5**	**6.8**

The best result is in bold.

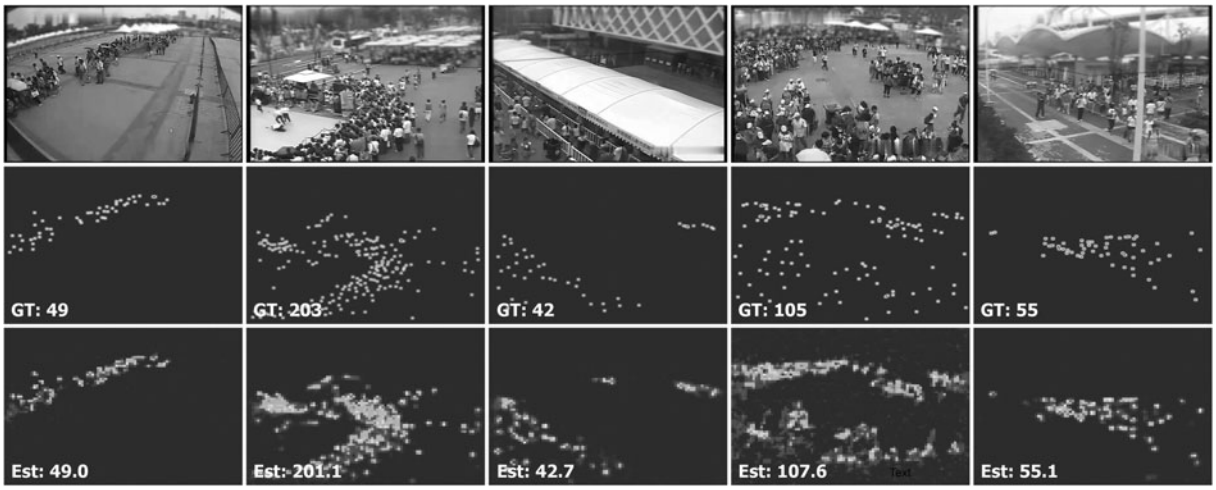

FIG. 8. Visually qualitative analysis of the WorldExpo'10 data set. The first, second, and third row show the input images, ground truth density maps, and estimated density maps, respectively.

the-art (SOTA) methods are shown in Table 4. It shows that the proposed SFANet performs best in Scenes 1, 3, and 5. Meanwhile, it achieves the best result in average MAE, with a reduction of 5.5% compared with the second-best method. Figure 8 shows some estimated results in the WorldExpo10 data set. One can see that the proposed method can accurately reflect the actual crowd distribution in all the images.

Performance on NWPU-Crowd data set. The NWPU-Crowd data set[39] is a new but challenging data set. It is composed of 5109 images with 2,133,375 head annotations, in which 3109 images for training, 500 images for validation, and 1500 images for testing. The data set has more challenges over other data sets such as high resolution, large density range (from 0 to 20,033), and mass negative samples. The quantitative results for NWPU-Crowd are listed in Table 5. It shows that the proposed SFANet scores 107.9 in MAE, which performs best, and 404.1 in RMSE, which ranks third, respectively. Especially, compared with Spatial/Channel-wise Attention Regression,[40] which also adopts attention mechanism, the proposed SFANet reduces the score of RMSE by 18.4%. The visualization of estimated maps with counting numbers is shown in Figure 9. It demonstrates that the proposed method performs well in the dense crowds with the accurate estimation.

Ablation study

To further verify the effectiveness of critical components, that is, BSA and MSCA proposed in SFANet, a series of ablation studies are conducted. The counterparts are denoted as follows:

1. "baseline" represents the basic model that only adopt ResNet-50.
2. "baseline+BSA" refers to the addition of the BSA module to the "baseline."
3. "baseline+MSCA" denotes the addition of the MSCA module to the "baseline."
4. "baseline+MSCA_BSA" represents MSCA module is connected in series with BSA module as MSCA in front and BSA behind.
5. "baseline+MSCA||BSA" denotes that MSCA and BSA are added to baseline in parallel.
6. "baseline+BSA_MSCA" represents the proposed SFANet.

Table 5. Experimental results on the NWPU-Crowd data set

Methods	*MAE*	*RMSE*
TinyFaces[67]	272.4	764.9
MCNN[21]	232.5	714.6
SANet[68]	190.6	491.4
A-CCNN[47]	176.5	520.6
ADMG[69]	152.8	907.3
AutoScale[70]	122.6	468.3
CRSNet[15]	121.3	**378.8**
PCC-Net[32]	112.3	457.0
SCAR[40]	110.0	495.3
TransCrowd[71]	117.7	451.0
SFANet(ours)	**107.9**	404.1

The best result is in bold.
NWPU, Northwestern Polytechnical University.

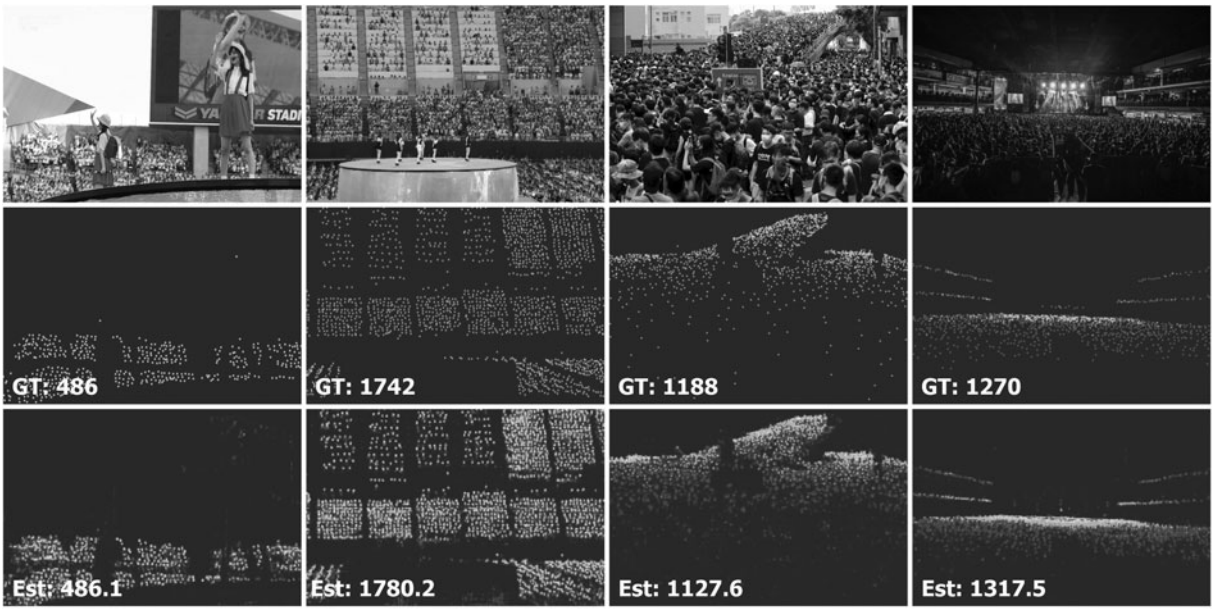

FIG. 9. Visually qualitative analysis of the NWPU-Crowd data set. The first, second, and third row show the input images, ground truth density maps, and estimated density maps, respectively. NWPU, Northwestern Polytechnical University.

Following the previous works,[15,41] we employ ShanghaiTech Part_A as the benchmark for the ablation study. The overall quantitative performance is shown in Table 6. It indicates that the two critical components, BSA module and MSCA module, contribute to the substantial improvement of the baseline method in terms of both MAE and RMSE. The MSCA module performs better in MAE (75.5) and RMSE (129.1) than that of the BSA module (76.2 and 133.0).

Analysis on connection mode, the "baseline+MSCA_BSA" module has a side effect. Both the "baseline+MSCA_BSA" module and the "baseline+BSA_MSCA" module enhance the baseline, but the latter is better. Theoretically, crowd counting can be divided into two steps. The first step is to select a discriminative area, and the second step is to count the number of heads in this area. BSA module completes the first step, and MSCA module completes the second step. Therefore, the latter is better than the former. The final SFANet boosts the baseline significantly by 5.9% and 5.0% in terms of MAE and RMSE, respectively.

Table 6. Ablation analysis of the key components in SFANet on ShanghaiTech Part A data set

Methods	*MAE*	*RMSE*
Baseline	77.9	134.4
baseline+BSA	76.2	133.0
baseline+MSCA	75.5	129.1
baseline+MSCA_BSA	105.6	169.1
baseline+MSCA‖BSA	75.2	128.6
baseline+BSA_MSCA	**71.7**	**122.5**

The best result is in bold.
BSA, bottleneck spatial attention; MSCA, multispectral channel attention.

Visualization of analysis results are demonstrated in Figure 10. The BSA module guarantees the accurate location of heads, as depicted in the red box in Figure 10d. The MSCA module can alleviate the error estimation for background regions, as depicted in the green box in Figure 10e. Different connection mode shows different performances. The "baseline+MSCA_BSA" mode presents unsatisfactory results, as depicted in the yellow box in Figure 10f. Owing to the reverse sequence of the two modules, the identification area is wrongly selected. The "baseline+MSCA‖BSA" (Fig. 10g) and "baseline+BSA_MSCA" (Fig. 10h) mode boost the estimation accuracy, the former results in the final overestimate due to overfitting, with the latter being more effective.

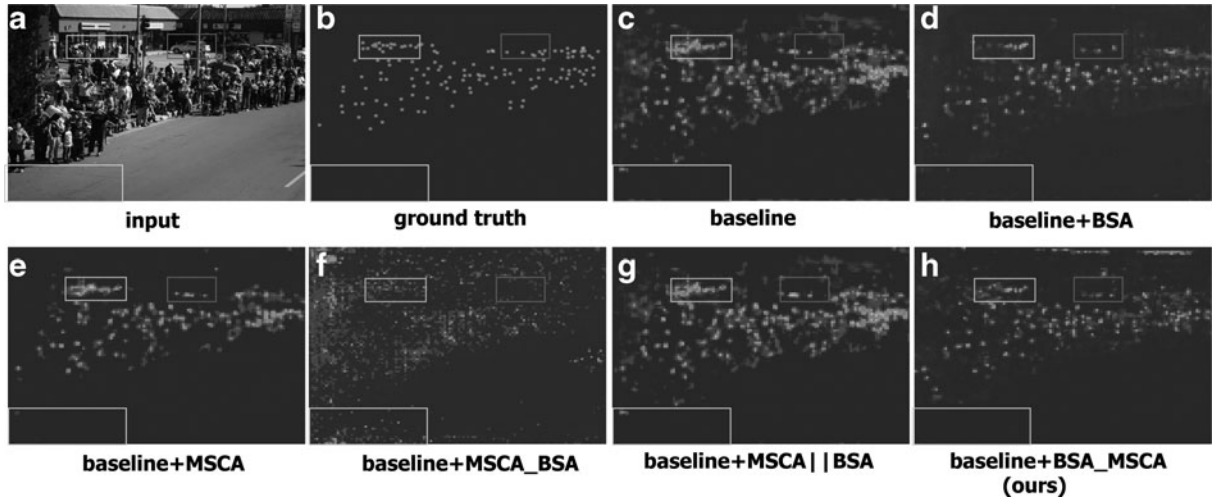

FIG. 10. The qualitative comparison of the baseline with different components. The white box depicts the dense crowds in high-density scenes, and the grey box illustrates the background cluster.

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 propose a SFANet, which consists of a BSA module and an MSCA module. The BSA module is built in the spatial domain to guarantee the accurate location of heads, whereas the MSCA module is built in frequency to accurately identify a person with multiple frequency components. These two attention modules highlight the crucial information in spatial and channel spaces in a mutual-promotion manner. Comprehensive experiments on five benchmark data sets prove that the SFANet achieves compelling performance on accuracy and robustness compared with the SOTA methods.

Acknowledgment

Thanks to Professor Qinqin Wei for her helps in revising the article.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

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

References

1. Liu W, Salzmann M, Fua P. Context-aware crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019. pp. 5094–5103. [Epub ahead of print]; DOI: 10.1109/CVPR.2019.00524.
2. Jiang J, Wang X, Gao M, et al. Abnormal behavior detection using streak flow acceleration. Appl Intell. 2022. [Epub ahead of print]; DOI: 10.1109/WiSPNET45539.2019.9032845.
3. Chan AB, Liang ZSJ, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008. pp. 1–7. [Epub ahead of print]; DOI: 10.1109/CVPR.2008.4587569.
4. Ren S, He K, Girshick RB, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2015;39:1137–1149.
5. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. 2015. pp. 1440–1448. [Epub ahead of print]; DOI: 10.1109/ICCV.2015.169
6. Idrees H, Saleemi I, Seibert C, Shah M. Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013. pp. 2547–2554. [Epub ahead of print]; DOI: 10.1109/CVPR.2013.329.
7. Gao G, Gao J, Liu Q, et al. CNN-based density estimation and crowd counting: A survey. 2020; arXiv abs/2003.12783.
8. Sindagi V, Patel V. A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognit Lett. 2018;107:3–16.
9. Qin Z, Zhang P, Wu F, Li X. FcaNet: Frequency channel attention networks. In: Proceedings of the International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021. pp. 783–792.
10. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 779–788. [Epub ahead of print]; DOI: 10.1109/CVPR.2016.91.
11. Amirgholipour S, Jia W, Liu L, et al. Pdanet: Pyramid density-aware attention based network for accurate crowd counting. Neurocomputing. 2021;451:215–230.
12. Rong L, Li C. Coarse- and fine-grained attention network with background-aware loss for crowd density map estimation. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV). 2021. pp. 3674–3683. [Epub ahead of print]; DOI: 10.1109/WACV48630.2021.00372.

13. Zhai W, Li Q, Zhou Y, et al. DA2net: A dual attention-aware network for robust crowd counting. Multimed Syst. 2022. [Epub ahead of print]; DOI: 10.1007/s00530-021-00877-4.
14. 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. [Epub ahead of print]; DOI: 10.1145/3343031.3350881.
15. Li Y, Zhang X, Chen D. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 1091–1100. [Epub ahead of print]; DOI: 10.1109/CVPR.2018.00120.
16. Ahmed I, Anisetti M, Jeon G. An IoT-based human detection system for complex industrial environment with deep learning architectures and transfer learning. Int J Intell Syst. 2021. [Epub ahead of print]; DOI: 10.1002/INT.22472.
17. 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: ICPR 2008 19th International Conference on Pattern Recognition. Tampa, FL: IEEE Computer Society, 2008. pp. 1–4.
18. Ge W, Collins RT. Marked point processes for crowd counting. In: CVPR. 2009. [Epub ahead of print]; DOI: 10.1109/CVPR.2009.5206621.
19. Lempitsky V, Zisserman A. Learning to count objects in images. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A (Eds.): Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2010. pp. 1324–1332.
20. Chan AB, Vasconcelos N. Bayesian poisson regression for crowd counting. In: 2009 IEEE 12th International Conference on Computer Vision. 2009. pp. 545–551. [Epub ahead of print]; DOI: 10.1109/ICCV.2009.5459191.
21. Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. pp. 589–597. [Epub ahead of print]; DOI: 10.1109/CVPR.2016.70.
22. Sam DB, Surya S, Babu RV. Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 4031–4039. [Epub ahead of print]; DOI: 10.1109/CVPR.2017.429.
23. Cheng J, Xiong H, Cao Z, Lu H. Decoupled two-stage crowd counting and beyond. IEEE Trans Image Process. 2021;30:2862–2875.
24. de Santana Correia A, Colombini E. Attention, please! a survey of neural attention models in deep learning. 2021; arXiv abs/2103.16775.
25. Liu N, Long Y, Zou C, et al. ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 3225–3234. [Epub ahead of print]; DOI: 10.1109/CVPR.2019.00334.
26. Sindagi VA, Patel VM. HA-CCN: Hierarchical attention-based crowd counting network. IEEE Trans Image Process. 2020;29:323–335.
27. Jiang X, Zhang L, Xu M, et al. Attention scaling for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 4705–4714. [Epub ahead of print]; DOI: 10.1109/cvpr42600.2020.00476.
28. Park J, Woo S, Lee JY, Kweon IS. BAM: Bottleneck attention module. 2018; arXiv:abs/1807.06514.
29. Gao J, Lin W, Zhao B, et al. c^3 framework: An open-source pytorch code for crowd counting. 2019; arXiv abs/1907.02724.
30. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014; arXiv: 1412.6980.
31. Sindagi V, Patel V. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2017. pp. 1–6. [Epub ahead of print]; DOI: 10.1109/AVSS.2017.8078491.
32. Gao J, Wang Q, Li X. PCC Net: Perspective crowd counting via spatial convolutional network. IEEE Trans Circuits Syst Video Technol. 2020;30: 3486–3498.
33. Zhang L, Shi Z, Cheng MM, et al. Nonlinear regression via deep negative correlation learning. IEEE Trans Pattern Anal Mach Intell. 2021;43:982–998.
34. Zhang Y, Zhou C, Chang F, Kot A. Multi-resolution attention convolutional neural network for crowd counting. Neurocomputing. 2019;329:144–152.
35. Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 532–546. [Epub ahead of print]; DOI: 10.1007/978-3-030-01216-8_33.
36. Zhang C, Li H, Wang X, Yang X. Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. pp. 833–841. [Epub ahead of print]; DOI: 10.1109/CVPR.2015.7298684.
37. Wang Q, Han T, Gao J, Yuan Y. Neuron linear transformation: Modeling the domain shift for crowd counting. IEEE Trans Neural Netw Learn Syst 2021:1–13. [Epub ahead of print]; DOI: 10.1109/TNNLS.2021.3051371.
38. Yan Z, Zhang R, Zhang H, et al. Crowd counting via perspective-guided fractional-dilation convolution. IEEE Trans Multimed. 2021. [Epub ahead of print]; DOI: 10.1109/TMM.2021.3086709.
39. 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. 2021;43:2141–2149.
40. Gao J, Wang Q, Yuan Y. SCAR: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing. 2019;363:1–8.
41. Bai S, He Z, Qiao Y, et al. Adaptive dilated network with self-correction supervision for counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 4593–4602. [Epub ahead of print]; DOI: 10.1109/cvpr42600.2020.00465.
42. Sindagi VA, Yasarla R, Babu DS, et al. Learning to count in the crowd from limited labeled data. In: Proceedings of the European Conference on Computer Vision (ECCV). 2020. pp. 212–229. [Epub ahead of print]; DOI: 10.1007/978-3-030-58621-8_13.
43. Sam DB, Babu RV. Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA: AAAI, 2018. pp. 7323–7330.
44. Sindagi V, Patel V. Generating high-quality crowd density maps using contextual pyramid CNNS. In: Proceedings of the International Conference on Computer Vision (ICCV). 2017. pp. 1879–1888. [Epub ahead of print]; DOI: 10.1109/ICCV.2017.206.
45. Huang S, Li X, Zhang Z, et al. Body structure aware deep crowd counting. IEEE Trans Image Process. 2018;27:1049–1059.
46. Zhang L, Shi M, Chen Q. Crowd counting via scale-adaptive convolutional neural network. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV). 2018. pp. 1113–1121. [Epub ahead of print]; DOI: 10.1109/WACV.2018.00127.
47. Kasmani SA, He X, Jia W, et al. A-ccnn: Adaptive ccnn for density estimation and crowd counting. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). 2018. pp. 948–952. [Epub ahead of print]; DOI: 10.1109/ICIP.2018.8451399.
48. Lei Y, Liu Y, Zhang P, Liu L. Towards using count-level weak supervision for crowd counting. Pattern Recogn. 2021;109:107616.
49. Liu L, Jiang J, Jia W, et al. DENet: A universal network for counting crowd with varying densities and scales. IEEE Trans Multimedia. 2021;23:1060–1068.
50. Sam DB, Peri S, Sundararaman MN, et al. Locate, size and count: Accurately resolving people in dense crowds via detection. IEEE Trans Pattern Anal Mach Intell. 2021;43:2739–2751.
51. Oh M-H, Olsen P, Ramamurthy K. Crowd counting with decomposed uncertainty. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2020. pp. 11799–11806. [Epub ahead of print]; DOI: 10.1609/AAAI.V34I07.6852.
52. Ding X, Lin Z, He F, et al. A deeply-recursive convolutional network for crowd counting. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 1942–1946. [Epub ahead of print]; DOI: 10.1109/ICASSP.2018.8461772.
53. Ranjan V, Le HM, Hoai M. Iterative crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 278–293. [Epub ahead of print]; DOI: 10.1007/978-3-030-01234-2_17.
54. Wang B, Liu H, Samaras D, Hoai M. Distribution matching for crowd counting. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2020.
55. Hu Y, Jiang X, Liu X, et al. Nas-count: Counting-by-density with neural architecture search. In: Proceedings of the European Conference on Computer Vision (ECCV). 2020. pp. 747–766. [Epub ahead of print]; DOI: 10.1007/978-3-030-58542-6_45.

56. Liu X, Yang J, Ding W. Adaptive mixture regression network with local counting map for crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV). 2020. pp. 241–257. [Epub ahead of print]; DOI: 10.1007/978-3-030-58586-0_15.
57. Abousamra S, Hoai M, Samaras D, Chen C. Localization in the crowd with topological constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Online Virtual, AAAI, 2021. pp. 872–881.
58. Liu L, Lu H, Zou H, et al. Weighing counts: Sequential crowd counting by reinforcement learning. In: Proceedings of the European Conference on Computer Vision (ECCV). 2020. pp. 164–181. [Epub ahead of print]; DOI: 10.1007/978-3-030-58607-2_10.
59. Wang Q, Lin W, Gao J, Li X. Density-aware curriculum learning for crowd counting. IEEE Trans Cybern 2020:1–13. [Epub ahead of print]; DOI: 10.1109/tcyb.2020.3033428.
60. Shi X, Li X, Wu C, et al. A real-time deep network for crowd counting. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 2328–2332. [Epub ahead of print]; DOI: 10.1109/ICASSP40776.2020.9053780.
61. Zou Z, Cheng Y, Qu X, et al. Attend to count: Crowd counting with adaptive capacity multi-scale cnns. Neurocomputing. 2019;367:75–83.
62. Yang Y, Li G, Wu Z, et al. Reverse perspective network for perspective-aware object counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 4373–4382. [Epub ahead of print]; DOI: 10.1109/cvpr42600.2020.00443.
63. Liu Y, Wen Q, Chen H, et al. Crowd counting via cross-stage refinement networks. IEEE Trans Image Process. 2020;29:6800–6812.
64. Miao Y, Lin Z, Ding G, Han J. Shallow feature based dense attention network for crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 34 2020. pp. 11765–11772. [Epub ahead of print]; DOI: 10.1609/AAAI.V34I07.6848.
65. Zhang AR, Yue L, Shen J, et al. Attentional neural fields for crowd counting. In: Proceedings of the International Conference on Computer Vision (ICCV). 2019. pp. 5713–5722. [Epub ahead of print]; DOI: 10.1109/ICCV.2019.00581.
66. Chen J, Su W, Wang Z. Crowd counting with crowd attention convolutional neural network. Neurocomputing. 2020;382:210–220.
67. Hu P, Ramanan D. Finding tiny faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1522–1530. [Epub ahead of print]; DOI: 10.1109/CVPR.2017.166.
68. 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). 2018. [Epub ahead of print]; DOI: 10.1007/978-3-030-01228-1_45.
69. Wan J, Chan AB. Adaptive density map generation for crowd counting. In: Proceedings of the International Conference on Computer Vision (ICCV). 2019. pp. 1130–1139. [Epub ahead of print]; DOI: 10.1109/ICCV.2019.00122.
70. Xu C, Liang D, Xu Y, et al. Autoscale: Learning to scale for crowd counting. Int J Comput Vis. 2022. [Epub ahead of print]; DOI: 10.1007/s11263-021-01542-z.
71. Liang D, Chen X, Xu W, et al. TransCrowd: Weakly-supervised crowd counting with transformer. 2021; arXiv:abs/2104.09116.

Cite this article as: Guo X, Gao M, Zhai W, Shang J, Li Q (2022) Spatial-frequency attention network for crowd counting. *Big Data* 3:X, 1–13, DOI: 10.1089/big.2022.0039.

Abbreviations Used

BSA = bottleneck spatial attention
CNNs = Convolution Neural Networks
DCT = discrete cosine transform
GT = ground truth
MAE = mean absolute error
MCNN = multicolumn CNN
MSCA = multispectral channel attention
NWPU = Northwestern Polytechnical University
PCC = pespective crowd counting
RMSE = root-mean-square error
ROI = regions of interest
SFANet = spatial-frequency attention network
SOTA = state-of-the-art
UCF-QNRF = University of Central Florida-Qatar National Research Fund
UCF_CC_50 = University of Central Florida_Crowd Counting_50