SOFA-Net: Second-Order and First-order Attention Network for Crowd Counting

Haoran Duan
h.duan5@newcastle.ac.uk

Shidong Wang
shidong.wang@newcastle.ac.uk

Yu Guan
yu.guan@newcastle.ac.uk

Open Lab, School of Computing
Newcastle University
Newcastle Upon Tyne, UK

Abstract

Automated crowd counting from images/videos has attracted more attention in recent years because of its wide application in smart cities. But modelling the dense crowd heads is challenging and most of the existing works become less reliable. To obtain the appropriate crowd representation, in this work we proposed SOFA-Net(Second-Order and First-order Attention Network): second-order statistics were extracted to retain selectivity of the channel-wise spatial information for dense heads while first-order statistics, which can enhance the feature discrimination for the heads' areas, were used as complementary information. Via a multi-stream architecture, the proposed second/first-order statistics were learned and transformed into attention for robust representation refinement. We evaluated our method on four public datasets and the performance reached state-of-the-art on most of them. Extensive experiments were also conducted to study the components in the proposed SOFA-Net, and the results suggested the high-capability of second/first-order statistics on modelling crowd in challenging scenarios. To the best of our knowledge, we are the first work to explore the second/first-order statistics for crowd counting.

1 Introduction

Crowd counting aims to count the number of people in images or videos of crowd scenes. It plays a pivotal role in real-world applications such as video surveillance, traffic planning, public security, etc. Earlier attempts were based on pedestrian detection [21] or human segmentation [27] in crowd. Recently, crowd counting has been regarded as an image-based density map regression task, and counting can then be conducted through integrating the densities. Density-map based methods achieved promising counting results in crowded scenes when it's difficult to detect subjects due to distance, occlusions, etc.

Previous density map regression works [7, 15] were proposed to learn the regional objects mapping, and recently Convolution Neural Network (CNN) became the major technique for crowd representation learning. In [18], a switching CNN learning inherent structural and functional differences is proposed to tackle large scale and perspective variations in crowd counting. Due to the diverse number of subjects and the various dense or sparse crowd

patterns, most recent CNN-based approaches [1, 11, 12, 14, 24] were proposed to estimate density maps by handling the multi-scale problems in crowd scenes. However, these methods become less reliable when the areas of pedestrians' heads are dense and very small. In [3], it was found that high density areas tended to be underestimated, while the low density areas tended to be overestimated. This observation suggested that a better representation should be learned in such challenging crowd scenarios. So we proposed a deep Second-Order and First-order Attention Network (SOFA-Net) for crowd modelling. Second-order statistics learning was successfully used to improve the representation learning [2, 5, 10, 23] or to recognize small objects in remote sensing [2, 22]. In this work, the second-order statistics leads our model to learn robust crowd representation by retaining selectivity of spatial information. First-order statistics, which can capture the discriminated spatial characteristic for crowd, was also used as complementary information. A Statistic-Wise Convolution operation was also proposed to effectively transform the second/first-order statistics into attentions for our network. Then a deep attention architecture was designed to handle multiple feature streams for generating the crowd density maps. For better generation quality [12, 28], a normalization strategy and a scale enhancement were also used. Our main contribution can be summarized as:

- To the best of our knowledge, this is the first work proposed to use second/first-order statistics for crowd modelling. We analysis the effects of second/first-order statistics for crowd counting qualitatively and quantitatively. Then, the overall experimental results suggested their feasibility in challenging crowd scenarios.

- We designed a multi-stream architecture with a Statistic-Wise Convolution to learn the second/first-order statistical attentions for crowd density map generation. Also, several tailored components were also proposed and evaluated.

- We tested our method on four popular public datasets, and it reached the state-of-the-art performance on most challenging datasets.

2 Related Work

Crowd Counting Early methods were based on designing detection/segmentation algorithms [21, 27], yet these methods may be heavily affected by occlusions, making them less practical. In [9], density map estimation approach was first introduced, which aimed to identify the centre locations of the subjects to avoid the error-prone detection procedures. CNN-based methods were the main techniques for representation learning in crowds counting [12, 13, 20, 25]. A maximum-excess-over-pixel loss was proposed with regional feature pattern to utilize the spatial information to count people in different density levels [3]. In [14], Liu et al. proposed a structured feature enhancement module by conditional random fields with a dilated multi-scale structural similarity loss to adapt the scale variations.

Second/First Order Statistic In large scale CNN network, the global average or max pooling was normally set at the end (as the first-order pooling) to capture the image representation by first-order statistical summary[19]. Recently, second-order statistics were also explored for improving the representation learning ability in many computer vision tasks [2, 5, 10, 23]. Li et.al evaluated the effectiveness of second-order information for large scale visual recognition with a trainable matrix power normalized covariance pooling [10]. The combination of first-order and second-order information was also employed in a multi-level architecture CNN [5] to improve the image texture discrimination. Based on second-order statistics, a

recurrent transformer network was proposed [2] to learn transformation-invariant representation for remote sensing with great performance at recognizing small objects.

3 Methodology

3.1 Problem Statement

Following [5, 7, 8, 9, 26], we describe the crowd counting problem as follows:

Given a crowd image $\mathbf{I}$ with c pedestrians' heads $\mathbf{H} = \{\mathbf{h}_i \in \mathbb{W}^2\}_{i=1}^c$, where $\mathbf{h}_i$ is the x-y coordinate of the i_{th} center point of the subject's head. The (ground truth) density map can be constructed by c Gaussian function $\mathcal{N}$ over all the heads' pixel grids in image $\mathbf{I}$, such that the crowd counts can be calculated by the integrals of density map. The (ground truth) density map can be written as:

$$\mathbf{D}^{gt} = \sum_{\mathbf{h} \in \mathbf{H}} \mathcal{N}\left(p; \mu = \mathbf{h}, \sigma^2\right) \tag{1}$$

where $p \in \mathbf{I}$ denotes the image pixels and σ is a very small number (spanning a few pixels [9]). Based on $\mathbf{I}$, we can clearly see $c = \sum_{p \in \mathbf{I}} \mathbf{D}_p^{gt}$. At the inference stage, given model $\mathcal{F}$ and the query crowd image $\mathbf{I}'$, the density map $\mathbf{D}^{pr}$ and the corresponding counts c^{pr} can be calculated as follows:

$$\mathbf{D}^{pr} = \mathcal{F}(\mathbf{I}', \hat{\mathbf{W}}), \quad c^{pr} = \sum_{p \in \mathbf{I}} \mathbf{D}_p^{pr}, \tag{2}$$

where $\hat{\mathbf{W}}$ is the model parameters.

In previous works, CNN was the major technique for density map regression, yet these models tended to overestimate low density crowd or underestimate high density crowd [4]. To learn robust crowd feature, here we propose a deep attention network by exploring the second-order and first-order statistics, and aggregation of these two complementary information may be essential for reliable crowd density estimation. The structure of our method is shown in Fig. 1.

3.2 SOFA-Net: Second-Order and First-Order Attention Network

Fig. 1 shows the overall architecture of our SOFA-Net, which consists of crowd features learning part and the density map regression part.

The crowd feature learning part starts from a feature encoding backbone, where we use the first 13 layers in VGG16 network [19]. Our proposed second/first-order statistical attentions are computed on the VGG16 feature maps at the end of this backbone. Specifically, the second-order statistics can be calculated based on (a derived) covariance matrix, while the first-order statistics can be extracted directly. The second-order and first-order statistics can be learned and transformed into attentions by a proposed Statistic-Wise Convolution operation. The two attentions can then be multiplied by crowd feature maps (extracted from VGG16), respectively for the second/first-order based features.

Given the second/first-order based features as well as the VGG16 features, we can estimate the crowd density map based on the generation blocks containing the Bilinear Upsampling layer and the basic Convolution operations. Although the main goal is to learn the
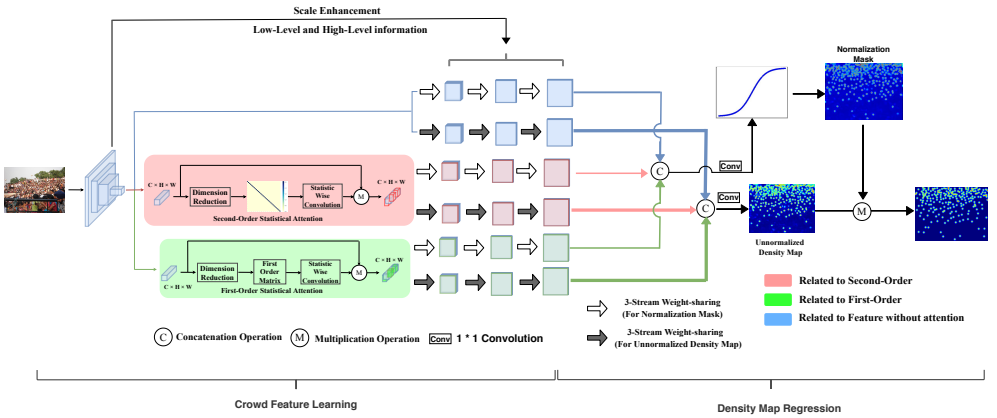
Figure 1: The overall framework of SOFA-Net. Pink colored components are related to second-order statistics; Green colored components are related to the first-order statistics, Blue colored components are related to the feature from VGG16 backbone.

second/first-order statistics for crowd representation, we also fuse the VGG16 backbone feature because of the summarized high-level semantic representation [19]. The aggregation of these three feature types may improve the quality of the generated density map. Moreover, a normalization mask is also devised and learned to normalize the density map for better quality. It is worth noting that the normalization mask and the (unnormalized) density map are learned separately, which may make both matrices less correlated for better normalization effect. Finally, motivated by the effectiveness of U-Net [17], we concatenate the output low-level and high-level crowd features from different backbone layers, which also leads model robust for scale variation. The entire network is optimized by a Pixel-Wise L_2 loss and a Position-Wise Binary Cross Entropy (BCE) loss [12]. The details of these components are given the next subsections.

3.2.1 Second-Order Statistical Attention

Recent works [2, 5, 11, 22, 23] suggested effective representation can be learned by second-order statistics for deep convolution neural network. Here we propose to use second-order statistics as an important component in our SOFA-Net (pink rectangle in Fig. 1), which guides the model to learn the channel-wise spatial information for crowd. Given extracted feature maps (height(h), width(w), channel(c)) from VGG16 backbone $\mathbf{F} \in \mathbb{R}^{h \times w \times c}$, 1×1 convolution, ReLU function and Batch Normalization are applied to reduce the channel dimension from c to c' obtaining $\mathbf{F}' \in \mathbb{R}^{h \times w \times c'}$. Then $\mathbf{F}'$ is flattened into $\mathbf{X} \in \mathbb{R}^{z \times c'}$ where $z = w \times h$. Covariance matrix $\mathbf{C}$ measuring the crowd correlation along channels can be formed as:

$$\mathbf{C} = \mathbf{X}\bar{\mathbf{I}}\mathbf{X}^T \tag{3}$$

where $\bar{\mathbf{I}} = \frac{1}{z}(\mathbf{I} - \frac{1}{z}\mathbf{1})$ with $\mathbf{I} \in \mathbb{R}^{c' \times c'}$ the identity matrix and $\mathbf{1} \in \mathbb{R}^{c' \times c'}$ is the matrix of all ones. After reshaping $\mathbf{C}$ to $\mathbf{C}' \in \mathbb{R}^{1 \times c' \times c'}$, the Statistic-Wise Convolution (see Fig. 2) is devised to normalize the covariance to obtain the inherent feature correlation and transform the second-order statistics into attention. The Statistic-Wise Convolution starts from a Independent
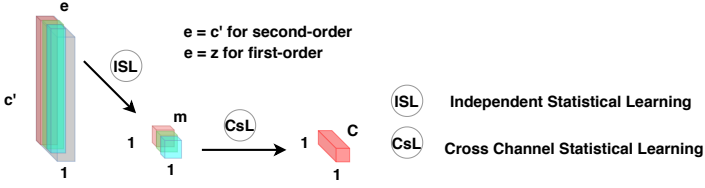
Figure 2: Statistic-Wise Convolution including two components: ISL/CsL.

Statistical Learning (see ISL component in Fig. 2) and the output feature maps $\mathbf{G} \in \mathbb{R}^{1 \times 1 \times m}$ can be formed as:

$$\mathbf{G}_{1,1,m} = \mathbf{K}_{1,c',m} \circ \mathbf{C}'_{1,c',c'} \tag{4}$$

where $\circ$ denotes the element-wise multiplication. The $m(=c')$ convolution kernels $\mathbf{K}$ are applied to the c' channels with the same size $c' \times 1$ of feature maps (vector in this case). This operation learns the statistical dependency along channels. Then a 1×1 convolution (i.e., CsL component in Fig. 2) is applied to learn the statistics with sharing convolution kernels and increase the dimension to c, so the second-order statistics are transformed into second-order attention as $\mathbf{G} \rightarrow \mathbf{A}_{so}$. This attention $\mathbf{A}_{so}$ is multiplied by the feature $\mathbf{F}$ (extracted from VGG16) to refine the representation. Specifically, the crowd feature with second-order attention is for med as $\mathbf{F}_{so} = \mathbf{F} \otimes \mathbf{A}_{so}$, where $\otimes$ is the multiplication operation between corresponding feature maps.

3.2.2 First-Order Statistical Attention

First-order statistics have been widely adopted in many CNN-based classification tasks to guide the back-propagation [16, 19]. The first-order statistics were also known to capture the spatial characteristic for texture in images [5]. Here we utilize the first-order statistics to preserve the subjects' head edges with the discrimination of heads and non-heads. Following similar feature extraction (from VGG16, i.e., with $\mathbf{F} \in \mathbb{R}^{h \times w \times c}$) and dimension reduction procedures (i.e., with $\mathbf{F}' \in \mathbb{R}^{h \times w \times c'}$), we have the feature matrix $\mathbf{X} \in \mathbb{R}^{z \times c'}$. Different from extracting second-order statistics, we directly use these spatial information. Specifically, after reshaping it into $\mathbf{X}_{fo} \in \mathbb{R}^{1 \times z \times c'}$, we learn the first-order attention $\mathbf{A}_{fo}$ by performing Statistic-Wise Convolution, i.e., $\mathbf{A}_{fo} = \mathbf{K}_{1 \times z \times c'} \circ \mathbf{X}_{fo}$. Similarly, the crowd feature with first-order attention is formed as $\mathbf{F}_{fo} = \mathbf{F} \otimes \mathbf{A}_{fo}$.

3.2.3 Density Map Estimation

The crowd density generation process are based on the aforementioned three feature types, i.e., crowd features $\mathbf{F}$ (e.g., extracted from VGG16), features with second-order statistical attention $\mathbf{F}_{so}$, and features with first-order statistical attention $\mathbf{F}_{fo}$. The final density map $\mathbf{D}^{pr}$ is generated based on two components, namely normalization mask $\mathbf{D}^{msk}$ and unnormalized density map $\mathbf{D}^{udm}$, as shown in Fig. 1. To reduce the correlation between these two components for better normalization effect, we train them through two-stream-like structure [12, 28].

A 3-stream weight-sharing scheme (via generation blocks f_G containing Bilinear Upsampling layer, 1×1 Convolution layer, 3×3 Convolution layer, Batch Normalization layer

and ReLU activation layer) is used for each component. The unnormalized density map $\mathbf{D}^{udm}$ is generated as:

$$\mathbf{D}^{udm} = \{f_G(\mathbf{F}, \mathbf{W}^{udm}), f_G(\mathbf{F}_{so}, \mathbf{W}^{udm}), f_G(\mathbf{F}_{fo}, \mathbf{W}^{udm})\}. \tag{5}$$

Note $\mathbf{W}_{udm}$ are the shared weights among the three feature streams. Similarly, the normalization mask can be calculated via $\mathbf{D}^{msk} = f_{sigmoid}(f_{conv}(\mathbf{D}'))$, where

$$\mathbf{D}' = \{f_G(\mathbf{F}, \mathbf{W}^{msk}), f_G(\mathbf{F}_{so}, \mathbf{W}^{msk}), f_G(\mathbf{F}_{fo}, \mathbf{W}^{msk})\}, \tag{6}$$

and $\mathbf{W}_{msk}$ are the shared weights among the three feature streams. The final density map can be then estimated through $\mathbf{D}^{pr} = \mathbf{D}^{udm} \circ \mathbf{D}^{msk}$.

For better training effect [28], we also applied the scale enhancement strategies, which concatenate the different scale information from different layers in backbone to density map generation block. The scale information allows the model to be more robust for heads' scale variation.

3.3   Optimization

To learn the model parameters, we use two loss functions in SOFA-Net: pixel-wise loss and position-wise loss.

Pixel-Wise Loss The Pixel-wise loss L_2 is defined as:

$$L_2 = \frac{1}{z} \sum_{i=1}^{W} \sum_{j=1}^{H} \left(\mathbf{D}_{i,j}^{gt} - \mathbf{D}_{i,j}^{pr} \right)^2 \tag{7}$$

where $\mathbf{D}^{pr}$ is the density map with height (H) and width (W) from the feed-forward operation and $z = W \times H$. It is the most widely used loss function on training deep convolution networks in crowd counting tasks [12, 14, 20].

Position-Wise Loss The position-wise loss L_{BCE} (Binary Cross Entropy) is calculated based on binarized ground truth $\mathbf{D}^b$, which comes from the $\mathbf{D}^{gt}$ based on a pre-defined threshold, and the (predicted) normalized density map (via Sigmoid) $\mathbf{S}^{pr} = f_{sigmoid}(\mathbf{D}^{pr})$ as follows:

$$L_{BCE} = -\frac{1}{z} \sum_{p=1}^{z} \left(\mathbf{D}_p^b \log\left(\mathbf{S}_p^{pr}\right) + \left(1 - \mathbf{D}_p^b\right) \log\left(1 - \mathbf{S}_p^{pr}\right) \right) \tag{8}$$

The Position-wise loss enforces the learning process to discriminate the crowd locations for better quality of density map generation [12]. The final loss function for SOFA-Net optimization is formulated as (in this work, we assign the lambda equal to 0.9).

$$L = \lambda L_2 + (1 - \lambda) L_{BCE} \tag{9}$$

4   Experiment

4.1   Implementation Details

Network Setting   The first 13 VGG16 layers (pre-trained model on ImageNet) were used to initialize the corresponding layers in SOFA-Net. Other parameters were initialized by

Table 1: Performance comparison on four public crowd counting datasets

Method	UCF_QNRF		ShanghaiTech A		ShanghaiTech B		UCF_CC_50	
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
[24]MCNN	-	-	110.2	173.2	26.4	41.3	377.6	509.1
[3]SCNN	-	-	90.4	135.0	21.6	33.4	318.1	439.2
[□]CSRNet	-	-	68.2	115.0	10.6	16.0	266.1	397.5
[□]RAZ-Net	116	195.0	65.1	106.7	8.40	14.1	-	-
[□]L2SM	104.7	173.6	64.2	98.40	7.20	11.1	188.4	315.3
[□]DSSINet	99.1	159.2	60.63	96.04	6.85	10.34	216.9	302.4
[□]MBTTBF	97.5	165.2	60.2	94.10	8.00	15.5	233.1	300.9
[5]SPANet	-	-	59.4	92.50	**6.50**	**9.9**	232.6	311.7
Ours	**96.2**	**158.7**	**57.5**	**92.12**	6.80	10.38	**185**	**281**

Gaussian disstribution with zero mean and 0.01 standard deviation. We set batch size to 50 and epoch number to 2000 in our experiments. We performed bilinear interpolation for any images less than 512×512, and the images were fed into network after randomly being cropped to 400×400 pixels. Observing that the images were collected from various situations of illumination, we adjusted the images by gamma contrast [0.5, 1.0] with the probability 25%. There are a few gray images in some datasets (e.g., ShanghaiTech_A), so in data augmentation we randomly converted a few (10%) for robust model training.

Datasets Our method was evaluated on the four popular public datasets, i.e., UCF_QNRF [8], ShanghaiTech [26](Part A and B), and UCF_CC_50 [7]. Out of them, UCF_QNRF contains large density variations and the subject number ranges from 49 to 12865. UCF_CC_50 includes extreme crowd scenes with serious noise. ShanghaiTech part A is very congested with noise, while ShanghaiTech part B is not congested. Following the protocol used in [12, 28], we generated the ground truth by a fixed Gaussian kernel. Also, the ground truth binary maps were generated by setting the threshold to 0.001 based on the ground truth density maps[12, 28]. For most train/test configurations, we followed the default protocols in the original papers (i.e., UCF_QNRF [8], UCF_CC_50 [7], ShanghaiTech_AB [26]). Notably, due to the limited sample numbers in UCF_CC_50, following [7] we set 5-fold cross validation for evaluation.

4.2 Evaluation Metrics

Following most existing works, the Mean Absolute Error (MAE) and Mean Square Error(MSE) were used as the evaluation metrics. Given predicted subject number c^{pr} (which can be inferred from $\mathbf{D}^{pr}$, see Eq. (2)), for N test crowd images the evaluation metrics can be calculated by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|c_i^{pr} - c_i^{gt}\right|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(c_i^{pr} - c_i^{gt}\right)^2} \qquad (10)$$

where c^{gt} denotes the ground truth heads' counting number.

4.3 Experimental Results

Model Comparison   Table 1 shows the results of our SOFA-Net and other state-of-the-arts on four afore-mentioned datasets. Our SOFA-Net outperforms others on most of the datasets (except ShanghaiTech Part B), which suggests its effectiveness on general crowd modelling tasks. Compared with the most recent works (i.e., methods in 2019) on UCF_QNRF dataset, SOFA-Net reaches much better results with further error reduction (i.e., in terms of MAE
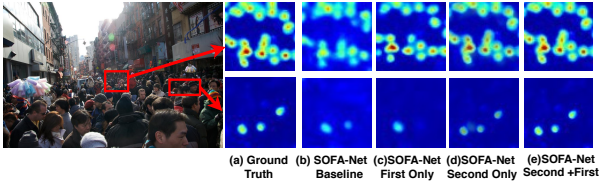
(a) Ground Truth (b) SOFA-Net Baseline (c)SOFA-Net First Only (d)SOFA-Net Second Only (e)SOFA-Net Second +First

Figure 3: The generated maps based on different settings of SOFA-Net in high density area (Top) and low density area (Bottom).

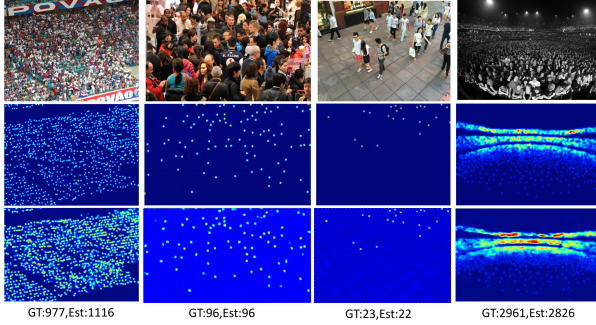GT:977,Est:1116   GT:96,Est:96   GT:23,Est:22   GT:2961,Est:2826

Figure 4: Some density maps generated by SOFA-Net; From top to bottom: original images, ground truth maps and generated maps

1.3 - 19.8 and MSE 0.5 - 36.3) than other methods. On ShanghaiTech Part A, our method is also much better in terms of both MAE and MSE. For ShanghaiTech Part B which was collected from shopping street with less crowded scenes, it can be seen that all the methods have good results in this relatively sparse and simple dataset. The performance can be further boosted by fusing other complementary information via the ensemble learning [6] or multi-stream structure [3, 4]. Nevertheless, our method outperforms most of the algorithms, and is comparable with state-of-the-art. UCF_CC_50 dataset, which includes very crowded scenes with high-levels of noises, was considered as the most challenging dataset. We can see our SOFA-Net can model the crowd counting tasks in these extreme conditions well, with much lower errors than other works (i.e., 3.4-234.5 in MAE and 19.9-260.6 in MSE).

Qualitative Analysis To understand better the effect of the proposed second/first-order statistical features, we visualized a challenging crowd image and the generated density maps (under different settings) in Fig. 3, from which some interesting observations can be made:

- From Fig. 3b, we can see without second/first-order statistical features (i.e., with feature $\mathbf{F}$ only), the generated crowd map is very blurry.

- With first-order statistical features (Fig. 3c, i.e., with features $\mathbf{F},\mathbf{F}_{fo}$), we can see clear boundaries among heads as enhanced discrimination, yet it cannot model the center-likelihood of dense heads areas well (with relatively low likelihoods in the centers).

- Retaining the selectivity of spatial information, features with second-order statistical attention (Fig. 3d, i.e., with features $\mathbf{F},\mathbf{F}_{so}$) can well preserve heads' areas (with high and precise likelihoods), which leads to the accurate counting. Finally aggregating both features can yield precise estimation (Fig. 3e, i.e., with features $\mathbf{F},\mathbf{F}_{fo},\mathbf{F}_{so}$).

From Fig. 3, we can clearly see second/first-order information are complementary for better crowd density map generation. We also generated several density maps in some challenging scenarios (as shown in Fig. 4), and results suggested its effectiveness even when there were more than thousand of people in crowd scenes.

4.4 Ablation Study

We also conducted ablation studies to quantitatively assess the core components in our SOFA-Net. ShanghaiTech part A, which covers various subject number in different scenes, was used as the benchmark dataset.

ShanghaiTech Part A dataset			
SOFA-Net	Features	MAE	MSE
No Attention	$\mathbf{F}$	68.6	109.3
First-Order	$\mathbf{F} + \mathbf{F}_{fo}$	65.6	104.2
Second-Order	$\mathbf{F} + \mathbf{F}_{so}$	60.8	97.1
Second/First-Order	$\mathbf{F}+\mathbf{F}_{fo}+\mathbf{F}_{so}$	**57.5**	**92.1**

Table 2: on the effect of second/first-order statistical attentions

ShanghaiTech Part A dataset		
SOFA-Net	MAE	MSE
Convolution	59.8	96.3
Statistic-Wise Convolution	**57.5**	**92.1**

Table 3: on the effect of Statistics-Wise Convolution

ShanghaiTech Part A dataset		
SOFA-Net	MAE	MSE
Without (Normalization) Mask	62.1	98.1
Mask (full-weight-sharing)	61.3	97.4
Mask (3-stream-weight-sharing)	**57.5**	**92.1**

Table 4: on the effect of normalisation masks

Effect on Second/First-Order Statistical Attention The core contribution of this work is the proposed second/first-order statistical attentions for robust representation learning. Fig. 3 demonstrated the effect of both components in a qualitative manner, and here we study them quantitatively. In Table 2, we report the SOFA-Net's results under different settings. With different attention types, we can use the corresponding feature combinations (e.g., $\mathbf{F},\mathbf{F}_{fo},\mathbf{F}_{so}$) to generate the density maps for crowd counting. We can clearly see that second-order statistical attention contributes the most to the performance, and the error rate can be further reduced if the complementary second/first-order statistical features were aggregated.

Effect on Statistic-Wise Convolution Statistic-Wise Convolution is a tailored operation that is proposed to learn second/first-order statistical attentions. The results in Table 3 suggests its effectiveness when compared with the standard convolution operation.

Effect on Normalization Mask In this work, we also trained a normalization mask to scale the generated density maps to avoid trivial results mostly in non-heads areas. In Table 4, we compared SOFA-Nets with/without normalization masks. For models with normalization masks, we also reported results with two different training strategies, i.e., the proposed 3-stream-weight-sharing scheme, as well as the full-weight-sharing scheme. Specifically, the former shared weights among the three feature streams $\mathbf{F},\mathbf{F}_{fo},\mathbf{F}_{so}$ to learn $\mathbf{W}^{udm}$ in Eq.(5) and $\mathbf{W}^{msk}$ in Eq.(6), respectively, while the latter shared weights among the three feature streams as well as the two tasks (with $\mathbf{W}^{udm} = \mathbf{W}^{msk}$). From Table 4, we can clearly see the normalization mask trained by the proposed 3-stream-weight-sharing scheme is much better than other structures. The result also suggests that a less correlated normalization mask (e.g., trained without weight-sharing between tasks) may further reduce the errors in such density map regression tasks.

5   Conclusion

In this work, we proposed SOFA-Net, which can extract second/first-order statistical attentions to learn robust representations for reliable crowd density map regression. The experimental results suggested second/first-order based features are complementary, and aggregating both features is feasible to reduce the error rate substantially for crowd density estimation. Also, the proposed method outperformed other state-of-the-arts in most (challenging) datasets. Although additional experiments and theoretical findings are necessary to draw the final conclusions on the benefit of applying second/first-order statistics for crowd counting, this work empirically demonstrates a simple yet effective way on modelling the crowd in challenging scenarios.

References

[1] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[2] Zan Chen, Shidong Wang, Xingsong Hou, Ling Shao, and Abu Dhabi. Recurrent transformer network for remote sensing scene categorisation. In *BMVC*, page 266, 2018.

[3] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G. Hauptmann. Learning spatial awareness to improve crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[4] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander G Hauptmann. Improving the learning of multi-column convolutional neural network for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1897–1906. ACM, 2019.

[5] Xiyang Dai, Joe Yue-Hei Ng, and Larry S. Davis. Fason: First and second order information fusion network for texture recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–28, 2017.

[7] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[8] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

[9] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332, 2010.

[10] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. *arXiv preprint arXiv:1807.00601*, 2018.

[14] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[15] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] Bin Qian, Jie Su, Zhenyu Wen, Devki Nandan Jha, Yinhao Li, Yu Guan, Deepak Puthal, Philip James, Renyu Yang, Albert Y Zomaya, et al. Orchestrating the development lifecycle of machine learning-based iot applications: A taxonomy and survey. *ACM Computing Surveys (CSUR)*, 2020.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[18] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE, 2017.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Vishwanath A. Sindagi and Vishal M. Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[21] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[22] Shidong Wang, Yu Guan, and Ling Shao. Multi-granularity canonical appearance pooling for remote sensing scene classification. *IEEE Transactions on Image Processing*, 29:5396–5407, 2020.

[23] Bryan (Ning) Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[24] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[25] Xin Zeng, Yunpeng Wu, Shizhe Hu, Ruobin Wang, and Yangdong Ye. Dspnet: deep scale purifier network for dense crowd counting. *Expert Systems with Applications*, 141:112977, 2020.

[26] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[27] Tao Zhao, Ram Nevatia, and Bo Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE transactions on pattern analysis and machine intelligence*, 30(7):1198–1211, 2008.

[28] Liang Zhu, Zhijian Zhao, Chao Lu, Yining Lin, Yao Peng, and Tangren Yao. Dual path multi-scale fusion networks with attention for crowd counting. *arXiv preprint arXiv:1902.01115*, 2019.