Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13664-8

1229: MULTIMEDIA DATA ANALYSIS FOR SMART CITY ENVIRONMENT SAFETY

Multiscale aggregation network via smooth inverse map for crowd counting

Xiangyu Guo[1] · Mingliang Gao[1] · Wenzhe Zhai[1] · Qilei Li[2] · Jinfeng Pan[1] · Guofeng Zou[1]

Received: 22 June 2022 / Revised: 3 August 2022 / Accepted: 11 August 2022

Abstract

Crowd counting is a practical yet essential research topic in computer vision, which has been beneficial to diverse applications in smart city environment safety. The commonly adopted paradigm in most existing methods is to regress a Gaussian density map that works as the learning objective during model training. However, given the unavoidable identity occlusion and scale variation in a crowd image, the corresponding Gaussian density map is degraded, failing to provide reliable supervision for optimization. To address this problem, we propose to replace the traditional Gaussian density map with a better alternation, namely the smooth inverse map (SIM). The proposed SIM can reflect the head location spatially and provide a smooth gradient to stabilize the model learning. Besides, we want the method to learn more discriminative features to cope with the challenge of large-scale variations. We deliver a multiscale aggregation (MA) to adaptively fuse features in different hierarchies to benefit semantic information under diverse receptive filed. The SIM and MA are meant to be complementary modules to guide the model in learning an accurate density map. Extensive experiments on benchmark datasets demonstrate the effectiveness of the proposed method compared with the state-of-the-art techniques.

Keywords Crowd counting · Smart city · Scale variation · Density map · Convolutional neural network

This article belongs to the Topical Collection: *1229: Multimedia Data Analysis for Smart City Environment Safety*
Guest Editors: Alessandro Bruno, Aladine Chetouani, Zoheir Sabeur, Marouane Tliba, Evangelos Maltezos, Miguel Gonzalez San Emeterio

✉ Mingliang Gao
mlgao@sdut.edu.cn

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

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

Published online: 24 August 2022

1 Introduction

The objective of crowd counting is to estimate the total number of heads in a given image or video sequence. The topic has drawn much attention due to its wide and practical application into smart city system, e.g., traffic control, and large gatherings [20]. Several crowd counting methods have emerged over the last years, along with the growth of the computer vision, especially with the advent of increasingly accurate deep learning models and architectures.

Early approaches to crowd counting fall into two categories, i.e., the detection-based and regression-based methods [2, 3, 6, 8, 27]. The detection-based methods usually rely on a sliding detector to count the people in a given image. Conversely, regression-based techniques directly map the number of heads from the image feature domain. Although these two approaches perform well in sparse scenes, they lose accuracy in densely crowded scenarios because of various challenges such as large-scale variation, background clutter, and severe occlusion, making the task more challenging.

The disruptive advent of deep learning fostered the design of several models for crowd counting. For instance, the convolution neural network (CNN)-based methods [4, 12, 33] have been adopted to address the problems above and achieved extraordinary results recently. The main idea behind the CNN-based method is to use a convolutional neural network to regress a density map, and the sum of pixel values on the density map estimates the number of people. The most commonly adopted density map in this domain is the Gaussian density map, which can provide a significant gradient for network training. While the density map provides both widespread and accurate training gradient, the network is still difficult to select the diverse scales and losses [17]. Ideally, a high-quality density map should provide a clear location of the heads in the image and provide continuous gradients for network training. Furthermore, large-scale variation has been the most common and challenging problem in crowd counting due to the irregular placement of the cameras. To handle this challenge better and enhance the counting accuracy, many multi-branch networks [1, 31, 33] have been proposed, which aims to extract the features of different sizes from each branch. Thus, generating a high-quality density map and integrating multiscale information are desired to enhance the counting performance.

In this paper, we introduce a novel method to generate the density map, namely the smooth inverse map. The primary advantages of the smooth inverse map can be divided into two aspects, i.e., the discriminative head locations and the beneficial gradient for network training. Figure 1 reveals the difference between the Gaussian density map and the proposed smooth inverse map. Figure 1(b) shows the Gaussian density map, which provides a set of Gaussian blobs and makes the head position blurred in congested scenes. Figure 1(c) is the smooth inverse map showing clear head positions in dense areas. Furthermore, we design the multiscale aggregation (MA) module to tackle the large-scale variation. In order to reduce the influence of scale variation, a natural way is to acquire diverse receptive fields to fuse spatial information of different sizes. Thus, the MA module adopts three branches with a skip connection to fuse features with diverse sizes. On this basis, we propose the multiscale aggregation network (MANet) for crowd counting. The MANet employs a feature extractor to extract the low-level feature. Then, we connect two multiscale aggregation modules to cope with the scale variation. And in the backend of the MANet, two transposed convolution layers are utilized to upsample the feature map and output the prediction. Overall, the contributions of this paper are summarized below.

1. A smooth inverse map is proposed for network training. The generation of the proposed map is based on distance transform, which is different from the primarily used Gaussian

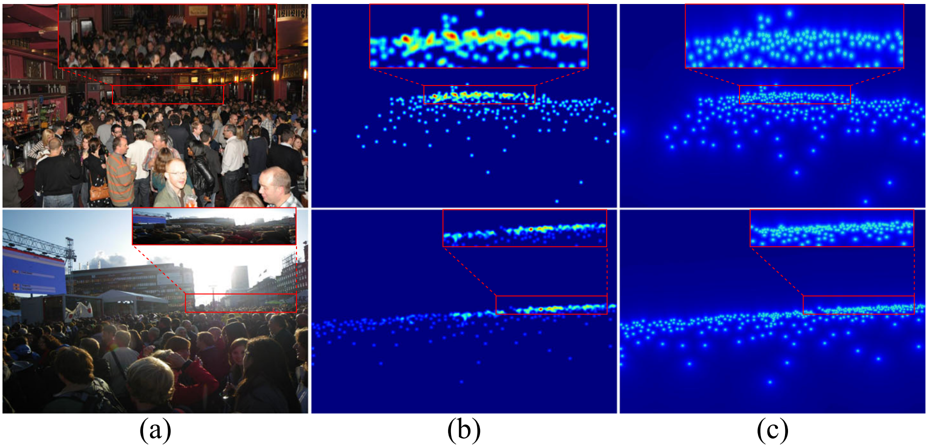

Fig. 1 Difference between smooth inverse map and Gaussian density map. The red bounding box denotes the congested crowd region. (a) Input image with large-scale variation and severe occlusions. (b) The Gaussian density map. (c) The smooth inverse map

density map. It can provide an accurate head location and a smooth gradient, which is helpful to guide the model to focus on the head region in training stage.

2. An MA module is designed to suppress the side effect of scale variations. Specifically, it builds a three-branch architecture, and each branch can obtain information at various scales with the help of different dilated convolution.
3. We perform comprehensive experiments on four crowd benchmark datasets to demonstrate the superiority of the proposed method in terms of counting performance.

2 Related work

The most related work to this work can be divided into two aspects, i.e., density map generation and multiscale feature fusion.

2.1 Density map generation

The generation of density maps plays a vanguard role for preparing the training data, and a high-quality density map is crucial in training the network. Zhang et al. [32] proposed a perspective information based density map, which aims to deal with the scale variation. Unfortunately, it is quite difficult to acquire a perspective map. Zhang et al. [33] figured out that the distance of two neighbouring persons could reflect the head size, and they introduced a geometry-adaptive kernel-based density map, which can tackle the scale variation. This geometry-adaptive map has been the most commonly adopted map for crowd counting. Taking into account the content-aware, Oghaz et al. [16] introduced the brute-force search algorithm and 2D Gaussian filters to acquire more content. Furthermore, Xu et al. [30] first employed the distance transform to generate a density map named distance labelled map, which can accurately show the position of the head and ensure there is no overlapping in dense areas. Olmschenk et al. [17] proposed the inverse kNN (ikNN) map, which is generated by utilizing a similar method to distance transformation. This map can provide a significant gradient for network training. Similar to [30], Liang et al. [13] also applied the distance transformation to generate density maps, which can represent exact person locations. The proposed density map can simultaneously deal with crowd localization and counting tasks. Based on the density map generation methods discussed above, we propose the smooth inverse map.

2.2 Multiscale feature fusion

The scale variation is still a challenging problem in crowd counting. A natural way to solve the scale variation is to fuse multiscale features. Zhang et al. [33] built a three-branch architecture to fuse the feature maps with different resolutions. Convolution kernels of different sizes (5,7, and 9) are utilized on each branch to acquire the multiscale features. The obtained features are fused through a convolution operation, and can better deal with the scale variation. Kasmani et al. [11] designed a model which can take advantages of contextual information with the local and global features to generate a high-quality density map. Specifically, the model is generated by choosing the best HyperParameters. This model can learn to tackle the scale variation in a simple way. Sindagi et al. [21] introduced a fusion scheme that can integrate information from different layers. The scheme is contributed to a scale-complementary block, which is adopted to acquire correlative features from the neighbouring layers. Gao et al. [5] proposed a network aiming to tackle the perspective changes and occlusions. It extracts features from three levels, i.e., global, local and pixel level, and then merge these features by a DULR (down, up, left and right) module. Sindagi et al. [22] employed the channel and spatial attention modules at different levels to integrate global and local information. The generated feature map contains rich spatial information, which is helpful to cope with the scale variation. Sajid et al. [18] utilized multi-layer branches to merge the information with each other by catering the crowd scale. Song et al. [24] designed a density head branch, to acquire density maps with different levels, then the feature maps are weighted to generate the estimated density map, which can relieve the scale variation well. Inspired by these methods, we propose the multiscale aggregation module to relieve the scale variation by fusing multiscale features.

3 The proposed method

3.1 Theoretical analysis of the density map generation

In this subsection, we analyse three state-of-the-art density maps, i.e., Gaussian density map, inverse kNN (ikNN) map and focal inverse distance transform (FIDT) map, which are most related to the proposed smooth inverse map.

Gaussian density map The Gaussian density map is the most commonly used density map. Supposing each head position is represented as a delta function. Then, a normalized Gaussian kernel is convolved with the delta function and generate the Gaussian density map. In a nutshell, it is formulated as,

$$M(\mathbf{z}, \sigma_k) = \sum_{i=1}^{H} \frac{1}{\sqrt{2\pi}\sigma_k} exp(-\frac{(\mathbf{z} - \mathbf{z_i})^2}{2\sigma_k^2}), \tag{1}$$

where $\mathbf{z}_i$ is a pair of coordinates representing the head position, and H represents the total number of head annotations in the image. σ_k denotes a value learned from other head locations. The value of σ_k is affected by the distance of the two nearest neighbour head. In dense regions, the heads are closely spaced, reflecting a good indication of head size. However, in sparse areas, the distance between the centre of the head annotation is large. The large σ_k may lead to the confusion of pedestrian position information on the density map. The distance distribution curve of the Gaussian map (σ_k=3) is marked in purple in Fig. 2(b). The distance distribution curve of the Gaussian density map tends to 0 faster than the other three maps, which results in the loss of distant training information. Meanwhile, the visualization of the Gaussian density diagram is provided in Fig. 2(c). It can be observed that the Gaussian blobs make the head's position indistinguishable.

Inverse kNN map To address the problem of Gaussian density map, the inverse kNN (ikNN) map [17] was proposed to retain more precise and distant training details. It is formulated as,

$$I = \frac{1}{M(\mathbf{z}, k) + 1}. \tag{2}$$

Here, $M(\mathbf{z}, k)$ is the full kNN map, which is formulated as,

$$M(\mathbf{z}, k) = \frac{1}{k}\sum \min_{k}(\sqrt{(\mathbf{z} - \mathbf{z_p})^2}, \forall p \in \mathcal{P}, \tag{3}$$

where $\mathcal{P}$ denotes a set of head positions. The distance distribution curve of ikNN map is shown in green in Fig. 2(b). It is observed that compared with the purple curve (Gaussian density map), the green curve provides a considerable and steep gradient of each person, which is beneficial to network training. In addition, compared with Fig. 2(c) and (d), the ikNN map can provide the discriminative head annotation location.

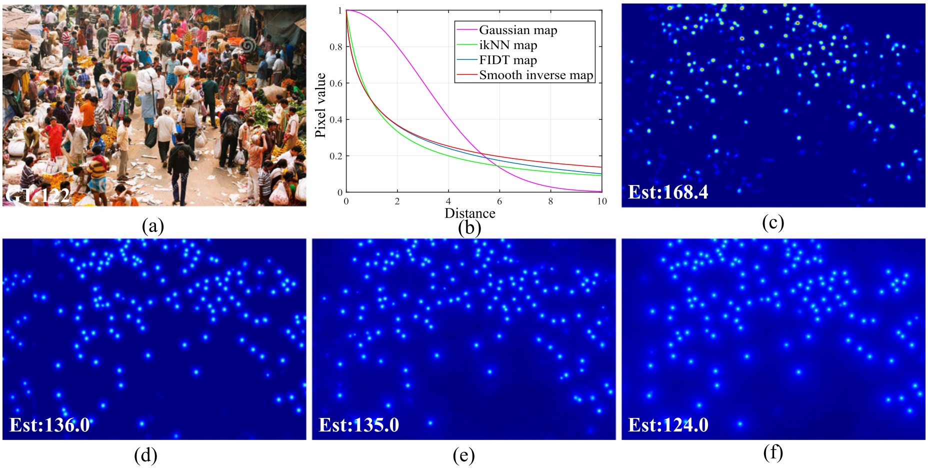

Fig. 2 The comparison of different density maps. The GT and Est denote the ground truth and estimated values, respectively. (a) Input image. (b) The distance from head annotation of Gaussian map, ikNN map, FIDT map and smooth inverse map. (c) Gaussian map. (d)ikNN map. (e) FIDT map. (f) smooth map

Particularly, when k equals to 1, the full kNN map has the same representation as the distance transform formulation. The formulation is provided as,

$$M(\mathbf{z}) = \min_{\mathbf{z_p}}(\sqrt{(\mathbf{z}-\mathbf{z_p})^2}), \forall p \in \mathcal{P}, \tag{4}$$

(4) represents the distance between the arbitrary pixel in an image and the nearest head annotation.

Focal inverse distance transform map The focal inverse distance transform (FIDT) map [13] refines the ikNN map ($k = 1$). It demonstrates that the ikNN map is a special case and can be optimized by fine-tuning the hyperparameters. In addition to crowd counting, the FIDT map can also carry out the task of crowd localization. Specifically, the FIDT map is defined as,

$$F = \frac{1}{M(\mathbf{z})^{(\alpha \times M(\mathbf{z})+\beta)} + 1}, \tag{5}$$

where α and β are two hyperparameters, and they are set as 0.02 and 0.75, respectively. (5) exhibits that the FIDT map adds a focal term, i.e., $\alpha \times M(\mathbf{z}) + \beta$, to the index part of the distance transform formulation, compared with the ikNN map. Based on the focal term, the FIDT map enables the model to concentrate on the head areas. In Fig. 2(b), the blue curve illustrates the distance distribution of the FIDT map. One can see that the pixel value of FIDT map decays slower than ikNN map in the head regions. Meanwhile, Fig. 2(e) shows that the FIDT map also ensures the head position while providing a curve that changes more slowly than the iKNN map. It means that the network should concentrate on the foregrounds [13].

3.2 Smooth inverse map

Based on the above theoretical analysis, a high-quality ground truth density map should contain the exact spatial location of the head. Beyond that, it is also necessary to provide continuous and useful gradients to make the density map focus as much on the head region as possible. Taking these points into account, we propose the smooth inverse map, which can provide a smoother gradient than other three maps while ensuring the precise position of each pedestrian. The proposed smooth inverse map is formulated as,

$$S = \frac{1}{M(\mathbf{z})^{(\alpha \times ln(1+M(\mathbf{z}))+\beta)} + 1}, \tag{6}$$

where $M(\mathbf{z})$ denotes the function of the distance transform. It's obvious to notice that we improve the distance transform formulation, aiming to make the pixel value decays slower in the head region than other maps. Specifically, we add a term, i.e., $\alpha \times ln(1 + M(\mathbf{z})) + \beta$ to the exponential part of the distance transform formulation. The improvement can make the density map provide a smoother curve (red curve in Fig. 2(b)), compared with the FIDT map. Hence, the map is named as smooth inverse map in this work. To ensure a fair comparison, the values of α and β remain the same as the configuration of FIDT map (0.02 and 0.75).

3.3 Multiscale aggregation module

The scale variation is a chronic problem in crowd counting. Most existing methods alleviate this problem by fusing multiscale information through multi-branch architecture. Meanwhile, multiply scales receptive fields are also desired because the size of heads varies continuously in an image. To this aim, we design a multiscale aggregation module which is composed of three branches and a skip connection. Figure 3 provides the architecture of the multiscale aggregation module.

Similar to [33], we adopt three branches to acquire various scale information. For each branch, a 1 × 1 convolution is firstly utilized to reduce the channels of the input feature map. Then, a 3 × 3 dilated convolution is adopted to enlarge the receptive fields. The dilated convolutions with different dilated ratios can capture more head sizes. Considering the head size is tiny, especially in dense crowds, the large dilated ratios may lead to large gap between the sizes of different receptive field. Furthermore, the crowd density is continuous, it requires a dense scale range. Therefore, we choose the dilated ratios as 1, 2, and 3 to better adapt the scale variation. Afterwards, we fuse the three refined features along channel dimension. In this way, the structure can acquire the multiscale information of the heads. Next, a skip connection is leveraged to concatenate the fused feature and the input feature. Finally, a 3 × 3 convolution layers is used to generate the estimated the results. The multiscale aggregation module can be represented as,

$$O(x) = x \oplus (\mathrm{Cat}(D^{i}\text{-Conv3}(\mathrm{Conv1}(x)))), i \in \{1, 2, 3\}, \tag{7}$$

where x and $O(x)$ denote the input and output features, respectively. $\oplus$ is a channel-wise summation operation. $D^{i}\text{-}Conv3$ represents the dilated convolution with dilated ration i.

3.4 The framework of the network

The architecture of the proposed multiscale aggregation network is depicted in Fig. 4. It consists of three modules, i.e., frontend network module, multiscale aggregation module and transposed convolution layer module.

We employ the HRNet [25] as frontend network to extract the basic feature map. The frontend network consists of four stages, and the output feature map is generated in the fourth stage. However, the extracted feature map suffers from the scale variation. To tackle this problem, we design the multiscale aggregation module. Specifically, two multiscale aggregation modules are cascaded after the HRNet to sufficiently integrate the multiscale information. Finally, two transposed convolution layers are utilized as the backend network to upsample the feature map to the same resolution as the input. The whole network can be formulated as,

$$P(I) = 2 \times \text{T-Conv3}(2 \times O_{s}(\mathrm{Cat}(\text{Stage-}i(I)))), i \in \{1, 2, 3, 4\}, \tag{8}$$

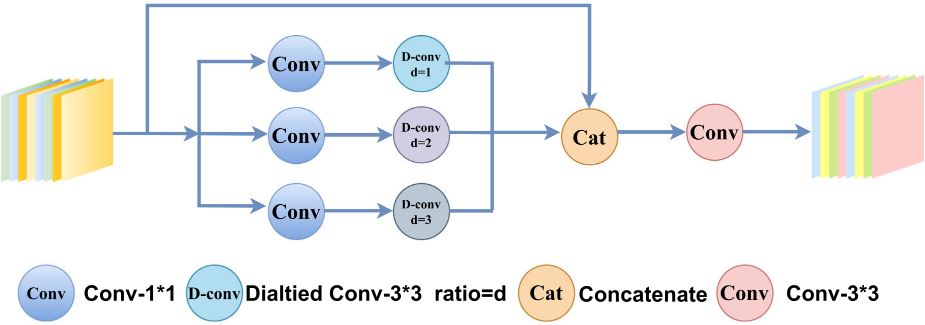

Fig. 3 The architecture of the multiscale aggregation module

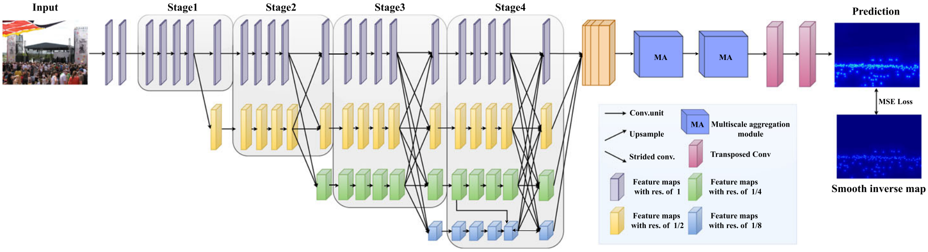
Input
Stage1
Stage2
Stage3
Stage4
MA
MA
Prediction
MSE Loss
Smooth inverse map
Conv.unit
Upsample
Strided conv.
MA Multiscale aggregation module
Transposed Conv
Feature maps with res. of 1
Feature maps with res. of 1/4
Feature maps with res. of 1/2
Feature maps with res. of 1/8

Fig. 4 The architecture of the proposed MANet for crowd counting

where I and P denote the input image and its prediction, respectively. $O_s(\cdot)$ represents a function of the multiscale aggregation module. Stage-i means the $i\text{-}th$ stage process of the HRNet. Cat$(\cdot)$ is the concatenated operation and T-conv$(\cdot)$ is the transposed convolution

3.5 Loss function

During training phase, we employ MSE loss to optimize the model by calculating the Euclidean distance between ground truth map and predicted density map. The loss function is defined as,

$$\mathcal{L}_{MSE}(E, G) = \frac{1}{N}\sum_{i=1}^{N}(E_i - G_i)^2, \tag{9}$$

where E and G denote the estimated value and the ground truth value, respectively. N is the total number of the heads.

4 Experiments and analysis

In this section, we evaluate the counting performance on four public crowd counting datasets, i.e., ShanghaiTech [33], UCF-QNRF [10], UCF_CC_50 [9] and JHU-Crowd++ [23]. Besides, the implementation details and evaluation metrics are detailed. Finally, a series of ablation experiments are conducted to prove the superiority of the components.

4.1 Implementation details

The training data is randomly cropped and horizontally flipped. The reason why not using vertical flipping is that the vertical flipping reverse the positions of the head and feet, which is unfit for counting accuracy. Specifically, for ShanghaiTech dataset, the crop size is set to 256 × 256. For other datasets, the crop size is set to 512 × 512. During the training stage, the batch size is set to 4 for UCF-QNRF dataset. For other datasets, the batch size is set to 8. We employ Adam optimizer with a learning rate at 1e-4 to train our network and the decay rate is set to 0.995. The experiments are conducted in the PyTorch framework [13] with two NVIDIA GTX3060 GPUs.

4.2 Evaluation metrics

To evaluate the performance of the proposed model, we apply the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), respectively.

$$MAE = \frac{1}{C}\sum_{i=1}^{C}\left|GT^{i} - Est^{i}\right|, \tag{10}$$

$$RMSE = \sqrt{\frac{1}{C}\sum_{i=1}^{C}\left|GT^{i} - Est^{i}\right|^{2}}, \tag{11}$$

where C denotes the number of testing samples. GT^{i} and Est^{i} represent the ground truth and estimated count of the i-th sample, respectively.

4.3 Comparative analysis

Comparative experiments are performed on four public crowd counting datasets, i.e., ShanghaiTech, JHU-Crowd++, UCF-QNRF and UCF_CC_50. The comparative results with the state-of-the-art competitors are presented in Table 1.

ShanghaiTech [33] is the most widely used crowd counting dataset. Due to the different attributes of the dataset, it is divided into two parts, i.e., PartA and Part B. The former consists of 482 images crawled from the websites, and it prefers high-density crowd scenes. By contrast, the Part B has 716 images collected from a busy street, and it focuses on sparse

Table 1 Comparative results on the Part A, Part B, JHU++, UCF-QNRF and UCF_CC_50 datasets

Method	Part A		Part B		JHU++		UCF-QNRF		UCF_CC_50	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [33]	110.2	173.2	26.4	41.3	188.9	483.4	277.0	426.0	377.6	509.1
SFCN [29]	64.8	107.5	7.6	13.0	77.5	297.6	102.0	171.4	214.2	318.2
A-CCNN [11]	85.4	124.6	19.2	31.5	171.2	453.1	367.3	-	-	-
LSC-CNN [19]	66.4	117.0	8.1	12.7	225.6	302.7	120.5	218.2	-	-
CSRNet [12]	68.2	115.0	10.6	16.0	85.9	309.2	-	-	266.1	397.5
PCCNet [5]	73.5	124.0	11.0	19.0	240.0	315.5	148.7	247.3	-	-
MUD-iKNN [17]	68.0	117.7	13.4	21.4	–	–	104.0	172.0	237.7	305.7
CG-DRCN [23]	64.0	98.4	8.5	14.4	71.0	278.6	112.2	176.3	–	–
SANet [1]	67.0	104.5	8.4	13.6	91.1	320.4	–	–	258.4	334.9
MBTTBF [21]	60.2	**94.1**	8.0	15.5	81.8	299.1	97.5	165.2	233.1	300.9
PaDNet [26]	59.2	98.1	8.1	**12.2**	–	–	96.5	170.2	185.8	278.3
KDMG [28]	63.8	99.2	7.8	12.7	69.7	268.3	105.6	180.5	–	–
RAZ [14]	65.1	106.7	8.4	14.1	–	–	116.0	195.0	–	–
HA-CCN [22]	62.9	94.9	8.1	13.4	–	–	118.1	180.4	256.2	348.4
DUBNet [7]	64.6	106.8	7.7	12.5	–	–	105.6	180.5	243.8	329.3
CAN [15]	62.3	100.0	7.8	**12.2**	100.1	314.0	107.0	183.0	212.2	301.3
Ours	**57.7**	101.9	**7.3**	14.3	**61.3**	**252.5**	**88.8**	**158.6**	**143.2**	**242.5**

The best results are highlighted in **bold**

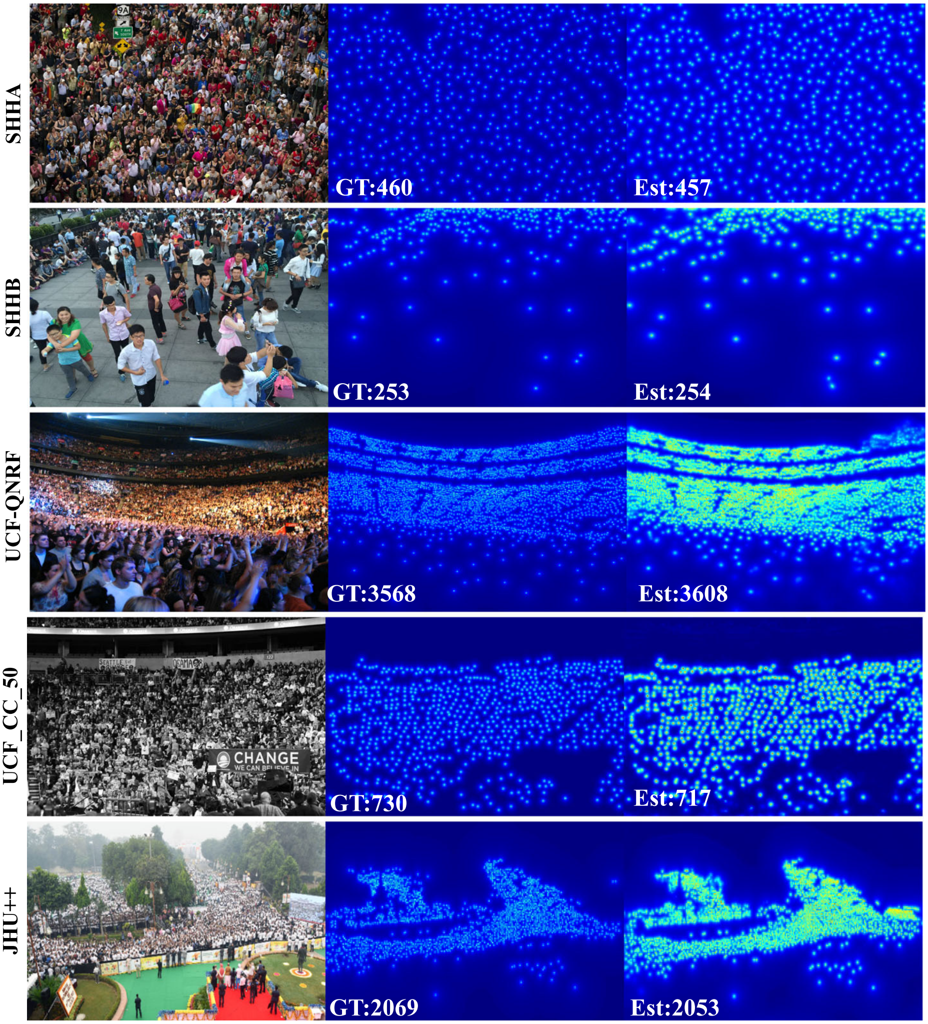

Fig. 5 Visualization results on different datasets. From tom to bottom are visualized results on Shanghai Part_A, Part_B, UCF-QNRF, UCF_CC_50 and JHU++ datasets. From left to right are the input image, the ground truth density map and the estimated density map

crowds. For Part A dataset, the MANet scores 57.7 in MAE, which ranks first place. Meanwhile, it scores 101.9 in RMSE, which is a competitive result. Specifically, compared with MUD-ikNN [17] which employs the inverse map, the proposed method improves the MAE and RMSE improved by 11.0% and 5.2%, respectively. For Part B dataset, the proposed model obtains the best result of 7.3 in MAE. Compared with PCCNet [5] which is specific to scale variation problem, the MANet improves the MAE by 33.6%. Furthermore, it achieves a competitive result in RMSE (14.3).

JHU-CROWD++ [23] is a dataset composed of 4,372 images (2,722 for training, 500 for validation, and 1,600 for test). The images present large scale without constraint. The head annotations are 1,515,005. The JHU-Crowd dataset also collects some weather-based images including rainy, hazy and snowy, which are crucial to train a robust model. It can be observed that the MANet exhibits the best results across the overall methods. Specifically, it scores 61.3 and 252.5 in MAE and RMSE, respectively. It improves the second-best method CG-DRCN [23] by 13.7% in MAE and 9.4% in RMSE. These results demonstrate that the MANet is fit for large-scale variation crowds.

UCF-QNRF [10] has 1,535 high-resolution images with wide viewpoints and varying lighting. There are 1.25 million head annotations on the dataset. It indicated that the UCF_QNRF is a fairly dense dataset. One can see that it achieves the best results of 88.8 and 158.6 in MAE and RMSE. Compared with the PaDNet [26], which designs a feature enhancement layer to fuse different scale information, while the MANet proposes the multiscale aggregation module to fuse multiscale information and outperforms it by 8.0% and 6.8% in terms of MAE and RMSE, respectively.

UCF_CC_50 [9] consists of 50 images, each presenting an extremely congested scene. The dataset has a total of 63,974 head annotations, with an average of 1,280 head annotations per image. It is a very challenging dataset due to the limited number of samples and the dense crowds in the image. On the dataset, it demonstrates that the MANet scores 143.2 and 242.5 which are both best in MAE and RMSE, respectively. Compared with the PaDNet [26], it has an improvement of 22.9% and 12.9%, respectively. Some visualization results on Shanghai Part_A, Part_B, UCF-QNRF, JHU++ and UCF_CC_50 datasets are shown in the Fig. 5. It proves that both the estimated crowd density map and counting number are closely related to the ground truth.

4.4 Ablation study

To further investigate the effectiveness of different components of MANet, we conduct ablation studies on Part A dataset. The ablation studies are divided into two aspects, i.e., the effectiveness of the smooth inverse map and the effectiveness of the multiscale aggregation module. The detailed network configurations are depicted as follows.

1. “baseline” represents the basic network that adopts HRNet with Gaussian density map.
2. “baseline+ikNN” represents the basic network that adopts HRNet with ikNN map.
3. “baseline+ FIDT” represents the basic network that adopts HRNet with FIDT map.
4. “baseline+ SI” represents the basic network that adopts HRNet with smooth inverse map.

Table 2 Ablation study on Part A

Methods	MAE	RMSE
baseline +kNN	69.3	138.7
baseline +ikNN	62.6	113.1
baseline +FIDT	62.1	**108.8**
baseline +SI	**60.3**	109.8
baseline +SI+SSM(1)	60.0	**100.0**
baseline +SI+SSM(2)	**57.7**	101.9
baseline +SI+SSM(3)	62.7	118.0

The best performances are highlighted in **bold**, and the second-best performances are underlined

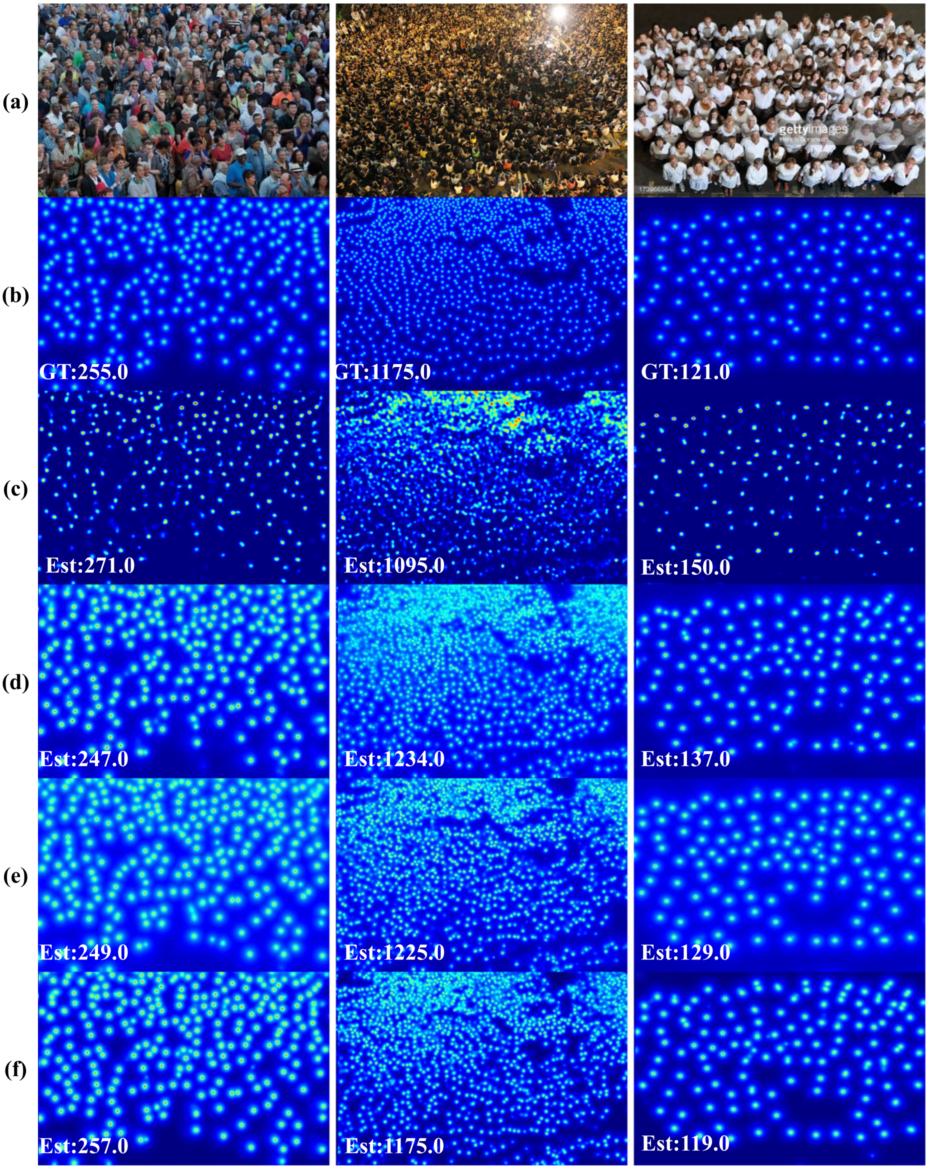

Fig. 6 Comparative results with different configurations. (a)∼(f) represent the input image, ground truth density map, estimated kNN map, estimated ikNN map, estimated FIDT map and estimated smooth inverse map. “GT” and “Est” denote the ground truth counts and the estimated counts, respectively

5. “baseline+ SI+ module(·)” represents the basic network that adopts HRNet with smooth inverse map, and add n module(s) to the HRNet.

1) Effectiveness of the smooth inverse map: We first explore the influence of the smooth inverse map. Table 2 shows the experimental results of different maps on ShanghaiTech Part_A. By comparing the first four rows, it can be seen that these four density maps perform well. The kNN map achieves the worst performance with the MAE and RMSE of 69.3 and 138.7. The ikNN map scores 62.6 and 116.1 in MAE and RMSE, respectively. Compared with the kNN map, it has an improvement by 9.7% and 18.5%, respectively. By contrast, the FIDT map performs better than the above maps. Its MAE and RMSE scores are 62.1 and 108.8, respectively. The proposed smooth inverse map exhibits the best MAE of 60.3. Although the score of RMSE is not the lowest, it is only 0.9% higher than the FIDT map. But the MAE is lower 2.9% than FIDT. The visually results are depicted in Fig. 6.

2) Effectiveness of the multiscale aggregation module: We figure out the impact of the multiscale aggregation module on counting performance. As reported in the last three rows in Table 2, one can observe that increasing a multiscale aggregation module is helpful to enhance the MAE and RMSE. Specifically, it achieves 0.5% and 8.9% improvement in MAE and RMSE, respectively. Increasing the number of multiscale aggregation modules to 2, the counting performance achieves the overall best performance with the MAE and RMSE of 57.7 and 101.9. Compared with the configuration without adding modules, the MAE and RMSE decrease 4.3% and 7.2%, respectively. Continue to increase the number of the multiscale aggregation modules to 3, and we can see a decrease in counting performance. Specifically, the scores of the MAE and RMSE increased by 8.7% and 15.8% compared with the best model.

5 Conclusion

In this paper, we propose a multiscale aggregation network with smooth inverse map for crowd counting. The proposed smooth inverse map is generated based on the distance transform. It can provide a smooth distance distribution curve, which is beneficial for regress a high-quality map. Experimental results indicate that the smooth inverse map performs better than other maps. The designed multiscale aggregation network adopts the HRNet as the backbone to extract the low-level feature, and two transposed convolution layers are utilized to upsample the density map. Two multiscale aggregation modules are employed to address the large-scale variation. The multiscale aggregation module is built with a three-branch architecture and a skip connection. It can alleviate the problem of scale variation by fusing the multiscale information. The multiscale aggregation network with the smooth inverse map achieves the state-of-the-art performance on four crowd counting datasets. In the future work, we will adopt the proposed smooth inverse map to crowd localization because it can provide distinguishable head position for each person in a congested region.

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

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

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate We declare that there is no ethics issue.

Conflict of Interests We declare that we have no conflict of interest.

References

1. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: ECCV. https://doi.org/10.1007/978-3-030-01228-1_45
2. Dollár P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34:743–761. https://doi.org/10.1109/TPAMI.2011.155
3. Felzenszwalb PF, Girshick RB, McAllester DA, Ramanan D (2009) Object detection with discriminatively trained part based models. IEEE Trans Pattern Anal Mach Intell 32:1627–1645. https://doi.org/10.1109/TPAMI.2009.167
4. Fu M, Xu P, Li X, Liu Q, Ye M, Zhu C (2015) Fast crowd density estimation with convolutional neural networks. Eng Appl Artif Intell 43:81–88. https://doi.org/10.1016/j.engappai.2015.04.006
5. Gao J, Wang Q, Li X (2020) Pcc net: Perspective crowd counting via spatial convolutional network. IEEE Trans Circuits Syst Video Technol 30:3486–3498. https://doi.org/10.1109/TCSVT.2019.2919139
6. Guo X, Gao M, Zhai W, Shang J, Li Q (2022) Spatial-frequency attention network for crowd counting Big data. https://doi.org/10.1089/big.2022.0039
7. Hwan Oh M, Olsen PA, Ramamurthy KN (2020) Crowd counting with decomposed uncertainty, arXiv:1903.07427. https://doi.org/10.1609/AAAI.V34I07.6852
8. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. In: 2013 IEEE Conference on computer vision and pattern recognition, pp 2547–2554. https://doi.org/10.1109/CVPR.2013.329
9. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2547–2554. https://doi.org/10.1109/CVPR.2013.329
10. Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the european conference on computer vision (ECCV), pp 532–546. https://doi.org/10.1007/978-3-030-01216-8_33
11. Kasmani SA, He X, Jia W, Wang D, Zeibots M (2018) A-ccnn: Adaptive ccnn for density estimation and crowd counting. In: 2018 25th IEEE International conference on image processing (ICIP), pp 948–952. https://doi.org/10.1109/ICIP.2018.8451399
12. Li Y, Zhang X, Chen D (2018) Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 1091–1100. https://doi.org/10.1109/CVPR.2018.00120
13. Liang D, Xu W, Zhu Y, Zhou Y (2021) Reciprocal distance transform maps for crowd counting and people localization in dense crowd. arXiv:2102.07925
14. Liu C, Weng X, Mu Y (2019) Recurrent attentive zooming for joint crowd counting and precise localization. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 1217–1226. https://doi.org/10.1109/CVPR.2019.00131
15. Liu W, Salzmann M, Fua PV (2019) Context-aware crowd counting. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 5094–5103. https://doi.org/10.1109/CVPR.2019.00524
16. Oghaz MM, Khadka AR, Argyriou V, Remagnino P (2019) Content-aware density map for crowd counting and density estimation. arXiv:1906.07258
17. Olmschenk G, Tang H, Zhu Z (2020) Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling, arXiv:1902.05379. https://doi.org/10.5220/0009156201850195
18. Sajid U, Ma W, Wang G (2021) Multi-resolution fusion and multi-scale input priors based crowd counting. In: 2020 25th International conference on pattern recognition (ICPR), pp 5790–5797. https://doi.org/10.1109/ICPR48806.2021.9412406
19. Sam DB, Peri SV, Sundararaman MN, Kamath A, Babu RV (2021) Locate, size, and count: Accurately resolving people in dense crowds via detection. IEEE Trans Pattern Anal Mach Intell 43:2739–2751. https://doi.org/10.1109/tpami.2020.2974830

20. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid cnns. In: 2017 IEEE International conference on computer vision (ICCV), pp 1879–1888. https://doi.org/10.1109/ICCV.2017.206
21. Sindagi VA, Patel VM (2019) Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: 2019 IEEE/CVF International conference on computer vision (ICCV), pp 1002–1012. https://doi.org/10.1109/ICCV.2019.00109
22. Sindagi VA, Patel VM (2020) Ha-ccn: Hierarchical attention-based crowd counting network. IEEE Trans Image Process 29:323–335. https://doi.org/10.1109/TIP.2019.2928634
23. Sindagi VA, Yasarla R, Patel VM (2022) Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE Trans Pattern Anal Mach Intell 44:2594–2609. https://doi.org/10.1109/tpami.2020.3035969
24. Song Q, Wang C, Wang Y, Tai Y, Wang C, Li J, Wu J, Ma J (2021) To choose or to fuse? scale selection for crowd counting. In: AAAI
25. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 5686–5696. https://doi.org/10.1109/CVPR.2019.00584
26. Tian Y, Lei Y, Zhang J, Wang JZ (2020) Padnet: Pan-density crowd counting. IEEE Trans Image Process 29:2714–2727. https://doi.org/10.1109/TIP.2019.2952083
27. Topkaya IS, Erdogan H, Porikli FM (2014) Counting people by clustering person detector outputs. In: 2014 11th IEEE International conference on advanced video and signal based surveillance (AVSS), pp 313–318. https://doi.org/10.1109/AVSS.2014.6918687
28. Wan J, Wang Q, Chan AB (2022) Kernel-based density map generation for dense object counting. IEEE Trans Pattern Anal Mach Intell 44:1357–1370. https://doi.org/10.1109/TPAMI.2020.3022878
29. Wang Q, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd counting in the wild. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 8190–8199. https://doi.org/10.1109/CVPR.2019.00839
30. Xu C, Liang D, Xu Y, Bai S, Zhan W, Tomizuka M, Bai X (2022) Autoscale: Learning to scale for crowd counting. Int J Comput Vis, pp 1–30. https://doi.org/10.1007/s11263-021-01542-z
31. Zhai W, Gao M, Anisetti M, Li Q, Jeon S, Pan J (2022) Group-split attention network for crowd counting, J Electron Imaging. https://doi.org/10.1117/1.JEI.31.4.041214
32. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks
33. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 589–597. https://doi.org/10.1109/CVPR.2016.70

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.