Multimedia Tools and Applications (2024) 83:61697–61714
https://doi.org/10.1007/s11042-023-14330-3

1229: MULTIMEDIA DATA ANALYSIS FOR SMART CITY ENVIRONMENT SAFETY

GTL-ASENet: global to local adaptive spatial encoder network for crowd counting

Chengming Liu[1] · Guanzhong Hu[1] · Yinghao Li[1] · Yufei Gao[1] · Lei Shi[1]

Received: 12 August 2022 / Revised: 9 November 2022 / Accepted: 2 January 2023 /
Published online: 3 March 2023

Abstract
Crowd counting from a single image is a challenging task due to perspective distortion and large-scale variation in crowd scenes. Many Researches only focus on local features to create density maps which is not effective in handing the challenges. This paper proposes a novel network named global-to-local adaptive spatial encoder network, which focuses on global features to generate a total structure density map of the population distribution, and then utilizes local features to reconstruct the total structure density map in detail to generate high-quality density map. To capture global features, local information and correlate them, we design a contextual module using different kernels with convolution and transposed convolution. To create a density map from global structure to local detail, two branches are designed, the global distribution branch and the local detail branch. The former aims to capture the population distribution region of interest in terms of global structure, and the latter aims to focus on the local details of each unit. Furthermore, to overcome the problem of pixel-wise loss of MSE, this paper proposes an efficient loss function that focuses on perceiving the possible crowd distribution over the whole image. We also apply a new upsampling mechanism that learns to create high-quality density maps on its own is advisable. The proposed network can capture the characteristics of pedestrian distribution and predict accurate results. It is evaluated on four crowd counting datasets (ShanghaiTech, NWPU, UCF_QNRF, UCF_CC_50), it obtains MAE of 67.1 and MSE, and achieves 108.8 in ShanghaiTech and gets MAE of 139.2 and the best MSE of 217.7 in UCF_CC_50 dataset and so on, and our method shows state-of-the-art on all the datasets.

Keywords Crowd counting · Density map · Spatial encoder · Global distribution · Contextual module

1 Introduction

Crowd counting which predicts the number of people from images/videos is important to many applications such as urban planning, public safety, space management and so on. But

✉ Chengming Liu
cmliu@zzu.edu.cn

Extended author information available on the last page of the article.

it's a challenging task owing to a large number of features similarity, perspective distortion and large scale variation.

Early heuristic crowd counting models fall into two categories: detection-based methods and regression-based methods. The former designs a sliding window to scan the entire image and detect pedestrians [8, 16, 17, 32]. Detection-based methods cannot handle scenes with large scale variation and occlusions between individuals. Then many regression-based algorithms [4, 5, 13] appeared to solve the crowd counting problems. But the main problem with these methods is always ignoring global and spatial features like SIFT [23] and HOG [7]. With the advent of deep learning, CNN-based algorithms have achieved remarkable performance in crowd counting. Several methods [9, 33, 34, 36, 42] implement a Basic-CNN architecture to calculate crowd counting, and achieve better performance than traditional computer vision-based methods. But they can't effectively encode large scale variation and diversified crowd distribution in congested scenarios. To address the the problems, multi-column architectures [1, 2, 26, 27, 43] are proposed to capture multi-scale features. However, multi-column network architectures are difficult to encode large scale variation and perspective distortion as similar network architectures have most same parameters. And training multi-column network architectures is not easy.

Due to the shortcomings of multi-column network architectures, simpler but effective single-column network architectures are widely used for crowd counting [3, 10, 12, 15, 18, 20]. They show good performance in creating density maps because they focus more on extracting and processing features, but they also struggle to challenge large-scale variation and diverse population distribution. In addition, the single-column approach suffers from three disadvantages at least. First, these architectures focus on local information while ignoring global and contextual information. Second, bilinear interpolation or convolution upsampling operators often lead to poor statistic distribution of predictions. Finally, MSE in loss function only focuses on pixel-wise correlation and ignores global structure.

To address the above problems, feature similarity, perspective distortion and large scale variation, a novel global-to-local adaptive spatial encoder network which try to solve above problems by crowd distribution and is proposed and using contextual information, unlike current methods focus on creating density maps by local information leads to density map difficult to create in hard region, model of this paper first focus on utilizing global structural information to create crowd distribution maps, and then based on the crowd distribution maps integrating global information and local cell details to generate density maps. Compared with current methods, the novelty of this model is as follow. Firstly, the model not only focuses local details but utilizes distribution of crowd to create density maps, Secondly, it uses contextual information to solve problems caused by perspective change, to specific, the model focuses on simple objects' change which are adjacent hard objects to predict hard objects.

The architecture of our model is shown in Fig. 1, where the first innovation is the contextual module. The contextual module sits behind the backbone and is designed to capture and correlate local and global information. Next, two branches are designed, the global distribution branch and the local detail branch, where the global distribution branch aims to generate high-quality density maps from the global structure. The latter consists of an adaptive spatial encoder module and a content-aware upsampling mechanism. The adaptive spatial encoder module consists of deformable convolutional layers and spatial encoder layers, which play an important role in encoding large-scale changes and diverse crowd distributions in crowded scenes. To create a better statistical distribution of density maps similar to ground truth density maps, a content-aware upsampling mechanism is introduced.

The main contributions of this paper are summarized as follows:

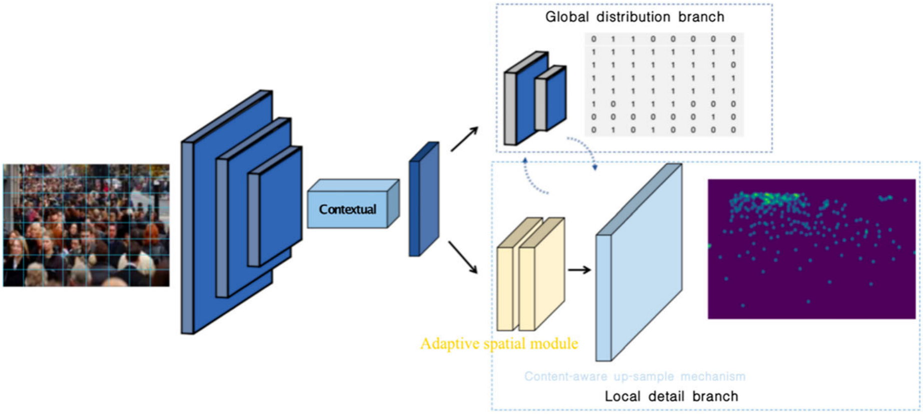

Fig. 1 Architecture overview of GTL-ASENet

(1) A novel Global-to-Local Adaptive Spatial Encoder Network (GTL-ASENet) is proposed, which can generate high-quality density maps from global structure to local details.

(2) Deep contextual information can be understand by contextual module.

(3) An adaptive spatial encoder module designed to adapt to complex and varied scenes, highlights useful crowd features, encodes complex geometric transformations and diverse crowd distributions.

(4) This paper deploys a content-aware upsampling mechanism that efficiently learns to cast feature maps to density maps.

The rest of this paper includes four sections. This paper first reviews the development process of crowd counting in Section 2. Section 3 introduces each module of the proposed method in detail, as well as the motivation and problems solved the module. Then, in Section 4, we introduce the evaluation criteria for crowd counting and the public datasets used, and give a detailed introduction to the performance of the model on every dataset. Meanwhile, this paper also analyzes results and the setting of the parameters. In the last section, we deduce the conclusions, meanwhile the future work and limitations of proposed study are given.

2 Related work

In this section, This paper reviews related works about crowd counting from basic-CNN, multi-scale models, local information models.

2.1 CNN-based models

This class of models uses the basic CNN architecture to estimate density maps and compute crowds without additional feature processing blocks. The first CNN-based method was proposed by Fu et al. [9], which designed a cascaded architecture to improve processing speed and prediction accuracy. Wang et al. [34] used the Alexnet architecture as a base and added many negative samples for counting. The CNN-based architecture is easy to apply, but usually not very accurate compared to state-of-the-art methods.

2.2 Multi-scale models

Several methods employ multiple branches to capture features at different scales, such as MCNN [43], Switch-CNN [1] and ACSCP [27]. MCNN proposes a multi-column architecture, where different branches use different convolution kernels to accommodate different receptive field features. Given that each branch needs to process a corresponding density, Switch-CNN adds a classifier to select the best branch to process image patches on a multi-column architecture. ACSCP uses an adversarial loss and splits the image into sub-blocks and parent blocks across scales to improve the performance of generating density maps. SASNet [30] introduces a bottom-up pyramid architecture designed to capture low-level and high-level features. To balance parameters and effectiveness, some methods choose VGG [29] or ResNet [11] as the backbone, such as SCAR [10] and SFCN [36]. Obviously, multi-column architectures have come a long way, but they still have some drawbacks. First, they are difficult to train because each branch needs to be trained individually. Furthermore, the number and density of crowds in the real world vary widely, it is difficult to design the number of branches. Finally, the bottom-up pyramid architecture consumes too much memory.

2.3 Local information models

The hallmark of such models is that they usually design elaborate encoders, such as adding attention mechanisms, introducing excellent upsampling operators. SANet [3] uses the inception module to capture multi-scale features, which consists of two parts: FME and DME, where FME introduces a scale aggregation module to address the independence between columns in MCNN [43]. DME is used to generate high quality density maps. SCAR [10] introduces spatial attention and channel attention to challenge the perspective changes of crowd scenes, and solves the dependence on the channel dimension through learning to improve the accuracy of regression. ADCrowdNet [20] introduces an attention map generator and a density map generator, where the former is used to develop the attention map, and the latter connects the input image and the output of the attention map generator to generate high-quality density maps. SAAN [12] applies an attention mechanism to fused density maps. All the above methods have good performance in generating density maps, but they only focus on local information and ignore the use of global information to make density maps.

3 The proposed method

3.1 Overview

In this section, flowchart of our networks is shown and contextual and adaptive spatial encoder module are described.

Our model mainly consists of two parts, an encoder and a decoder. The encoder is backbone for extracting efficient feature maps, and the decoder consists of two branches, the global distribution branch and the local detail branch. The global distribution branch is used to generate efficient crowd distribution maps that help the model understand the density map from the entire structure, the local detail branch aims to focus on globally distributed unit details. Specifically, first, an image is fed into the extractor, which uses ResNet-101

[11] as the backbone. Then, the characteristics are captured and enlarged with kernels of different sizes by a contextual models inspired by the Dilation module [41]. The output of the contextual module consists of 4 parts with 16-dimensional channels, which are connected in the channel dimension. It achieves significant improvements in the accuracy of mapping images to density maps, but still struggles with diverse crowd distributions in crowded scenes and distortions caused by perspective views. To this end, we use a global distribution branch to handle features that generate possible distributions in the global structure of crowd scenes, and an adaptation module is used to adapts to distortion by deformable convolutions that take the offset of sampling locations as learning parameters and population distribution. After this, a spatial encoder module is adopted to encode spatial features. Finally, the GTL-ASENet generates a 1-channel density map through the content-aware upsampling mechanism. For training , MSE (standard mean squared error) and BCE (binary cross-entropy loss) are used as loss functions.

3.2 Contextual module

Some researches such as CSRNet [18], SFCN [36] enable the model to obtain more spatial information through dilated convolution, and dilated convolution can increase the receptive field to obtain more spatial information, but it ignores the relationship between adjacent features. This module utilizes larger kernels to enlarge the receptive field and correlate local and global features, as shown in Fig. 2, which is an architecture consisting of convolution and transposed convolution. Specifically, this method applies 7 × 7, 5 × 5, 3 × 3 convolutions and 7 × 7, 5 × 5, 3 × 3 transposed convolutions to capture the feature maps of different effective feature sizes. The contextual model inherits the advantages of dilated convolution and extracts sufficient effective information, while it significantly avoids noisy information without padding. Our goal is to extract local and global features and correlate them with the contextual model, so it is essential and useful to use various kernels and concatenate the input and output features of the contextual module.

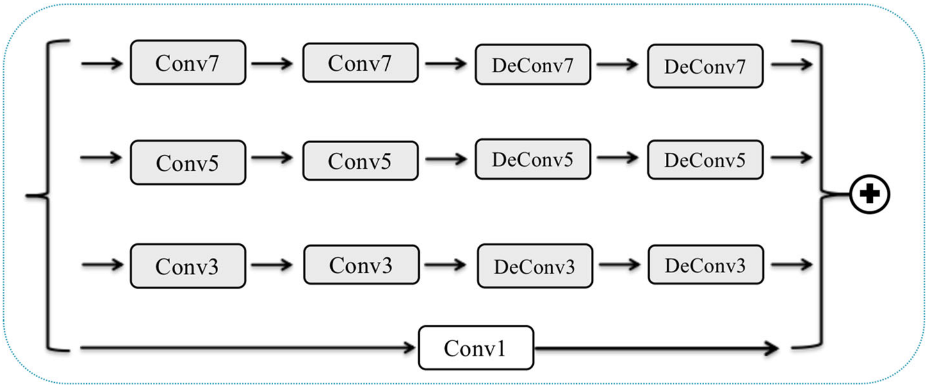

Fig. 2 The contextual module of GTL-ASENet

3.3 Adaptive spatial encoder module

In the scene of large-scale crowds, as the visual distance becomes farther, the objects become smaller and it is difficult to distinguish the objects from the background. However, the data distribution, the change of visual distance, the characteristics of pedestrians are similar in a certain area, so this paper focuses on the easy samples with short visual distance first, and uses the easy samples to predict the slightly difficult samples, and then uses the easy samples and the slightly difficult samples to predict the hard samples. Meanwhile, the distribution of pedestrians in different scenes is random, and the characteristics of pedestrians also change greatly with the increase of visual distance, which makes it difficult for the model to capture the characteristics and distribution of pedestrians. To challenge the tiny hard objects and diverse crowd distributions in crowded scenes, an adaptive spatial encoder module is designed, which consists of an adaptive module and a spatial encoder mechanism. The spatial encoder mechanism can deal with the random distribution of pedestrians in diverse scenes, and better perceive the law of crowd distribution. The adaptive module is used to solve the problem of huge changes in pedestrian characteristics in the same scene, and to grasp the law of pedestrian characteristics changes. The former use the simple objects to predict the difficult objects, while the latter understands the law of pedestrian distribution in areas where pedestrian characteristics change continuously.

Given a convolution kernel at K sampling positions, let $w(p_n)$ denote the weight at the n-th position, and p_n denote the learnable offset at the n-th position, R denote the regular grid for sampling the input feature map x. setting $R = \{(-1, -1), (-1, 0), \cdots, (0, 1), (1, 1)\}$, and using a deformable convolution scheme similar as [6]. The 2D modulated deformable convolution is formulated as

$$y(p_0) = \sum_{p_n \epsilon R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_n, \tag{1}$$

where $x(p_0)$ denotes the features at location p_0 from x, $y(p_0)$ denotes the output feature maps at location p_0, p_n belongs to R denoting the pre-specified offset, and Δm_n is a modulation scalar.

Random population spatial distribution information is obtained by utilizing the spatial encoder mechanism. Let F be a feature map of size $C \times H \times W$, which is first processed into H slices and then processed by a convolutional layer with C kernels of size $C \times w$, where w is the kernel width. The output of the convolutional layer is added to the next slice to generate a new slice. New slices are processed in the same way until the last slice is updated. It can be expressed as:

$$F_{c,w}^{h}{}' = \begin{cases} F_{c,w}^{h}, & h = 1 \\ F_{c,w}^{h} + R\left(\sum\limits_{m,n} F_{m,w+n-1}^{h}{}' \cdot W_{m,n}^{h}\right), & h \geqslant 2 \end{cases}, \tag{2}$$

where $F_{c,w}^{h}$ is the input tensor, c denotes channel, h and w indices row and column respectively, and R is the ReLU activation function.

3.4 Global distribution branch

To understand crowd distribution and help the model create density maps from the global structure, the branch of global distribution is designed. Specifically, method first concatenates the contextual module to obtain the output T, then down-samples T to 1/16 of the original image size, and then modulates C by the Sigmoid function. The value in C indicates the likelihood of anyone being present in the area. The ground-truth labels for C are generated from the ground-truth density map. Using maxpooling to process the ground-truth density map to obtain the ground-truth label Dot, $Dot_{i,j}$ represents the ground-truth label on region (i, j), which is defined as:

$$Dot_{i,j} = \begin{cases} 1 & Dot_{i,j} > threshold \\ 0 & otherwise \end{cases}. \tag{3}$$

The global distribution branch is supervised by a binary cross entropy(BCE) loss function:

$$L(C_{i,j}, Dot_{i,j}) = Dot_{i,j} log(C_{i,j}) + (1 - Dot_{i,j})(1 - log(C_{i,j})), \tag{4}$$

where $C_{i,j}$ is the predict possibility of region (i, j).

3.5 Content-aware up-sampling mechanism

By visualizing the outputs of the current methods, it is found that the density maps generated by up sampling introduced in current many methods has defects on the performance of local details. Specifically, the pedestrians' features in density maps are a process of gradual changes in the circle from the inside to outside, but the changes of local features in the density maps generated by current methods are not. To specific, many current methods' upsampling operator is the bilinear interpolation algorithm. However, the output of bilinear interpolation is different from the Gaussian distribution of the valid area of the ground-truth density map, which is generated by the Gaussian kernel function. Further more, bilinear interpolation cannot capture rich density information because only sub-pixel neighborhoods are considered. Another method of upsampling is deconvolution [24]. Unfortunately, deconvolution is prone to "uneven overlap", putting more of the metaphorical paint in some places. Developing density maps from feature maps is not just linear interpolation, but content, contextual information, and spatial feature transformations. Therefore, a content-aware upsampling mechanism is essential to learn the above transformations to generate high-quality density maps. Therefore method tries to introduce a method which can consider every feature point and content of feature map. Thus, this paper believes that different upsampling kernel should be used by different input contents, and each feature point should use its own upsampling kernel, rather than all feature points sharing the up-sampling kernel. Thus introducing CARAFE [35] as our upsampling operator to learn the above transformation. Given a feature map F of size $C \times H \times W$ and the upsampling size of $k_{up} \times k_{up}$, the kernel prediction module consists of three parts. First F is compressed from C to C_m convolutional layer of size 1×1, the predicted upsampling kennel size is $\sigma H \times \sigma W \times k_{up} \times k_{up}$. Second, the $k_{encoder} \times k_{encoder}$ convolutional layer is used to predict the upsampling kernel, resulting in a shape of $\sigma H \times \sigma W \times k_{up} \times k_{up}$. Finally, the predicted kernels are normalized using the Softmax function. The content-aware reconstruction module aims to reconstruct the function using the above-predicted upsampling kernel. For each reorganization kernel W_{out}, the content-aware reorganization module will reorganize the features within the local region through a weighted sum function. For the output position L_{out} and the corresponding square region $R(F, k_{up})$ centered on $L = (i, j)$, the process is formulated as (5),

$$X'_{L_{out}} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} W_{out(m,n)} \cdot X_{(i+n,j+m)}, \tag{5}$$

where $r = \lfloor k_{up}/2 \rfloor$ and setting k_up = 2.

4 Experiments

This section first describes implementation details and then describe the evaluation metrics and datasets followed by a detailed ablation study to understand the effects of different components in the proposed counting network. Finally, comparing results of the proposed method against several state-of-the-art methods on 4 publicly available datasets (NWPU [37], ShanghaiTech [43], UCF_QNRF [14], UCF_CC_50 [13]).

4.1 Implementation details

In all experiments, Adam is used as the optimizer and the initial learning rate is set to 0.1. Weight decay is stetted by 0.005. The number of iterations depends on complexity and the count of images. Backbone is the first 23 convolutional layers of Resnet101. For CARAFE σ is 8. Label of Global Distribution Branch is smoothed, positive object equals 0.998, negative object is 0.002 and loss of this branch is BCELoss. Input size depends on datasets, as well as batch size. Output is evaluated by MAE and MSE which calculate every corresponding pixel. MAE and MSE are recognized evaluate tools thus it is essential to prove authority of method.

4.2 Evaluation metrics

MAE and MSE are used to evaluate density map.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid D_i^{pred} - D_i^{gt} \mid, \tag{6}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i^{pred} - D_i^{gt})^2}, \tag{7}$$

where N is the number of images in one test sequence, D_i^{gt} is the ground truth of density map, and D_i^{pred} is the final output of model.

4.3 Datasets

NWPU [37] NWPU is collected by Qi Wang et al. NWPU is randomly split into three parts, namely training, validation and test sets, containing 5109 images, in a total of 2133375 annotated heads. Compared with existing crowd counting datasets, it contains various illumination scene and the largest density range from 0 to 20033. It's also the largest from the perspective of image and instance level.

ShanghaiTech [43] This dataset is collected by ShanghaiTech University. This dataset consists of two parts: Part_A and Part_B. Part_A contains 482 images and Part_B includes 716 images. Images in Part_A almost are token in congested scenes and most of them are randomly downloaded from the Internet. While images in Part_B are token from streets in Shanghai.

UCF_QNRF [14] The UCSD dataset was acquired with a stationary camera mounted at an elevation, overlooking pedestrian walkways.There are 1,535 crowd images and 1.25 million

Table 1 Statistics of the five crowd counting datasets

Dataset	Number	Resolution	Total	Min	Ave	Max
NWPU	5109	2191 × 3209	2,133,375	0	418	20,033
UCF-QNRF	1535	2013 × 2902	1,251,642	49	815	12,865
UCF_CC_50	50	2101 × 2888	63,974	94	1279	4,543
ShanghaiTech part A	482	589 × 868	241,677	33	501	3,139
ShanghaiTech part B	716	768 × 1024	88,488	9	123	578

Table 2 Performance of adaptive spatial encoder and contextual module on NWPU

Method	MAE	MSE	R2
Res101	107.67	543.4	0.35
Res101+adaptive	80.4	428.9	0.37
Res101+contextual	89	487.7	0.36
Res101+adaptive+contextual	76.9	401.7	0.39

Table 3 Performance of global distribution branch on SHHB

Method	MAE	MSE	R2
Res101	7.8	13.4	0.35
Res101+global	7.2	11.8	0.42
VGG	9.29	14.07	0.37
VGG+global	9.13	13.0	0.41
CSRNet	9.73	14.68	0.41
CSRNet+global	8.63	13.21	0.45

Table 4 Performance of global distribution branch on NWPU

Method	MAE	MSE	R2
Res101	107.67	543.4	0.002
Res101+global	80.2	468.7	0.0024

head annotations in UCF_QNRF, and this dataset has a wide range of counts. This dataset is a challenging dataset as the diversified scenes, extremely congested scenarios.

UCF_CC_50 [13] UCF_CC_50 only has 50 annotated images collected from internet, in a total of 67974 annotated heads. As the tiny number of images, diversified scenes and large amounts of individuals, this dataset is a challenge for every method.

The statistics of above datasets are shown in Table 1, which includes the number of images in corresponding dataset, image's resolution, total number of annotated people, the minimum and maximum number of annotated people in image and the average number of annotated heads. And we use a graph to show distribution of number range on three datasets in Fig. 5.

4.4 Ablation study

In order to demonstrate the effects of the proposed method, many ablation studies on ShanghaiTech PartB dataset and NWPU dataset to validate the effect of the proposed methods are advisable. Firstly, using ShanghaiTech PartB dataset to confirm the effectiveness of proposed methods, and then the necessity of global distribution branch is confirmed by four ablation experiments.

Effectiveness of adaptive spatial encoder and contextual module on the NWPU dataset is verified first. The experiment is divided into four categories according to the module combination: Res101, Res101+Adaptive, Res101+Contextual, and Res01+ Adaptive + Contextual. Then proving the validity of the global distribution branch on both SHHB and NWPU datasets. The verification experiment on the SHHB data is divided into two steps.

Table 5 The comparison results on the Shanghai Tech dataset

	PartA		PartB	
Method	MAE	MSE	MAE	MSE
MCNN [43]	110.2	173.2	26.4	41.3
CSRNet [18]	68.2	115.0	10.6	16.0
SCAR [10]	66.3	114.1	9.5	15.2
BL [22]	62.8	101.8	7.7	12.7
LA-Batch [45]	65.8	103.6	8.6	13.6
SFCN [36]	64.8	107.5	7.6	13.0
DUBNet [25]	64.6	106.8	7,7	12.8
TransCrowd [19]	66.1	105.1	9.3	16.1
Ours	67.1	108.8	7.0	11.7

Table 6 The comparison results on the NWPU dataset

Method	MAE	MSE
MCNN [43]	218.53	700.6
SANet [3]	171.16	471.51
CSRNet [18]	104.89	433.48
SCAR [10]	81.57	397.92
BL [22]	93.64	470.38
CAN [21]	93.6	489.9
SFCN [36]	95.46	608.32
CCFD [28]	76.8	343
MARNet [38]	80.8	422.8
Ours	74.4	390.5

First, Res101, VGG, and CSRNet are used as the baselines, and then the global distribution branch is added in these three baselines. The results of effectiveness of adaptive spatial encoder and contextual module on the NWPU dataset are shown in Table 2. With Res101 as the baseline, the MAE and MSE are 107.67 and 543.4 respectively. Then using the adaptive spatial encoder. The module was applied to Res101 and we got MAE 80.4 and MSE 428.9. This module improved MAE by 25% and MSE by 21%. Then we only apply the contextual module to the Res101, and the MAE obtained was 89, and the MSE was 487.7, thus the MAE increased by 17%, and the MSE increased by 10.3%. Finally, the Contextual Module and the adaptive spatial encoder module are applied to the Res101, and we got MAE 76.9 and MSE 401.7.

To prove the validity of the global distribution branch, conduction of three sets of comparative experiments on the SHHB data, using VGG, Resnet101 and CSRNet as baselines respectively is performed. In order to ensure the fairness of the experiments, the three methods used the same hyperparameters such as batch size, optimizer, and loss function. The results are shown in Table 3, from which we can see the global distribution branch has a good improvement effect for all the three methods. For VGG, MAE and MSE increased by 1.7% and 7.6%, respectively, for Res101, MAE and MSE increased by 7.7% and 11.9%, respectively, and CSRNet's MAE and MSE increased by 11.3% and 10%, respectively. Lastly, we verify the effectiveness on NWPU with Res101 as the baseline and the results

Table 7 The comparison results on the UCF_QNRF dataset

Method	MAE	MSE
MCNN [43]	277	426
TEDNet [15]	113	188
RANet [44]	111	190
S-DCNet [39]	104.4	176.1
DUBNet [25]	105.6	180.5
CAN [21]	107	183
LA-Batch [45]	113	210
Ours	101.3	187.3

Table 8 The comparison results on the UCF_CC_50 dataset

Method	MAE	MSE
MCNN [43]	377.6	509.1
TEDNet [15]	249.4	354.5
RANet [44]	239.8	319.4
SPN+L2SM[40]	188.4	315.3
DUBNet [25]	243.8	329.3
CAN [21]	212.2	243.7
LA-Batch [45]	203.0	230.6
M-SFANet [31]	162.33	276.76
SASNet [30]	161.4	234.46
Ours	139.2	217.7

are shown in Table 4 with Res101 as the baseline, and MAE increased by 25.5%, and MSE increased by 13.7%.

4.5 Comparisons with state-of-the-arts

In this section, we compare the proposed model with state-of-the-art methods on three challenging datasets.

Results on ShanghaiTech. As shown in Table 5, in PartA, our method obtains MAE of 67.1, and achieves 108.8 in MSE. In terms of Shanghai PartB, our model is the best in MAE of 7.0. In addition, ours also obtains the MSE of 11.7 which is the first best.

Results on NWPU. The comparison results on NWPU is shown in Table 6, where we get that our model achieves the best MAE of 74.4 and the second MSE 390.5.

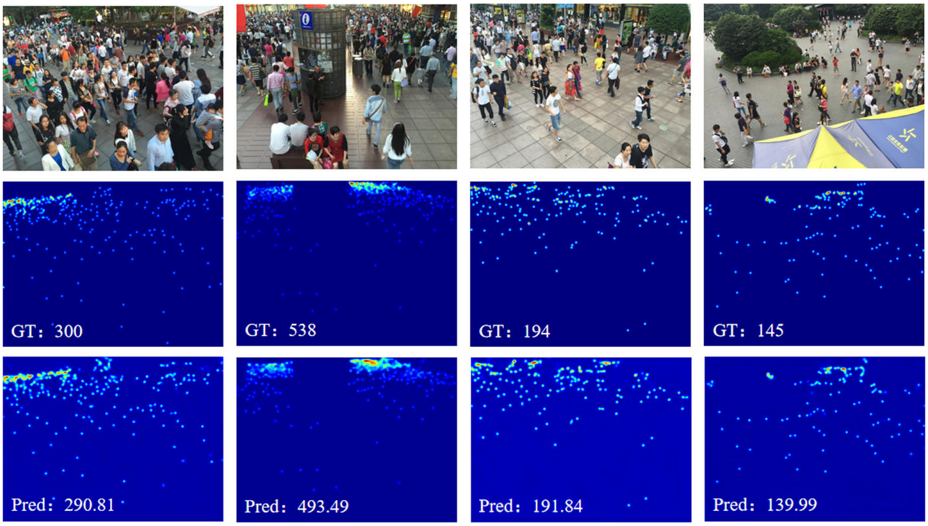

Fig. 3 Some density maps on ShanghaiTech. Row 1: original image, Row 2: groundtruth, Row 3: predicted density map by GTL-ASENet. "GT" denotes groundtruth count. "Pred" means the predict count

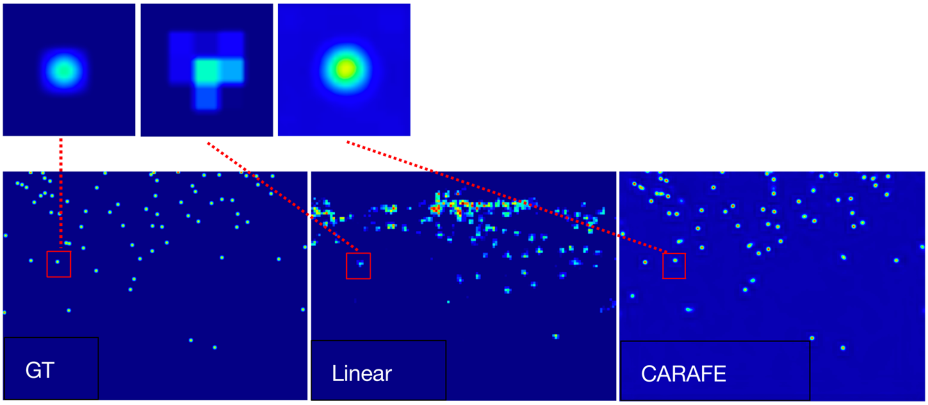

Fig. 4 The first row includes three region distribution at same position. From left to right, the first one is a unit in ground truth density map, the second one is a predicted unit with bilinear upsampling. The third is predicted by CARAFE upsampling. "GT" denotes groundtruth density map. "Linear" means using bilinear upsampling method. "CARAFE" means using CARAFE upsampling method

Results on UCF_QNRF. The comparison results of our method and other state-of-the-art methods on UCF_QNRF are shown in Table 7, our method obtains the best MAE of 101.3 which is better than S-DCNet by 3.0%.

Results on UCF_CC_50. As shown in Table 8, our model obtain the best MAE of 139.2 and the best MSE of 217.7. Compaerd with the second best of MAE and MSE, our method improves the MAE by 13.8% and the MSE by 5.6%.

4.6 Visualization results

Figures 3 and 4 are the visualization results generated by our GTL-ASENet. Figure 3 illustrates that the predict crowd distribution is very similar to the groundtruth and the estimation counting numbers are close to groundtruth counting numbers.

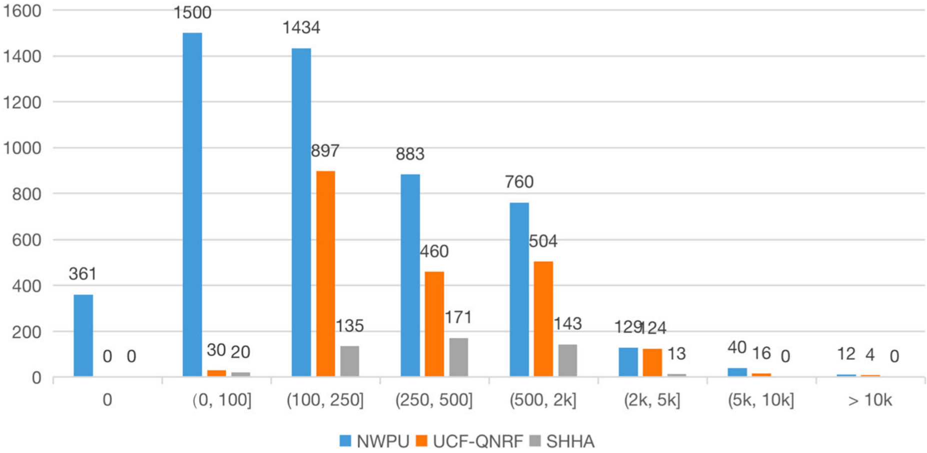

Fig. 5 The distribution of number range on three datasets

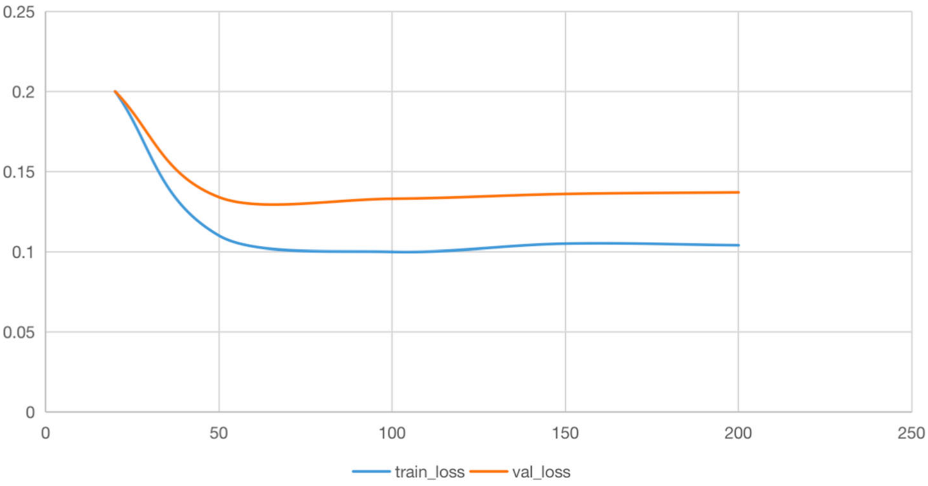

Fig. 6 The convergence speed graphs of train loss function and validation function

Figure 4 shows the comparison result of the bilinear upsampling method and the CARAFE, where the ground truth density map is created by a Gaussian kernel. The bilinear interpolation is not visible for mapping feature maps to density maps because the regions of interest in the bilinear interpolation output appear uniformly distributed. The CARAFE Fig. 5 method, on the other hand, outputs a better distribution, similar to the halo in the ground truth density map. Figure 6 shows the convergence speed graphs of train loss function and validation function, Fig. 7 shows the cure lines of MAE and MSE. Figure 8 shows the difference of ground truth and predicted density map

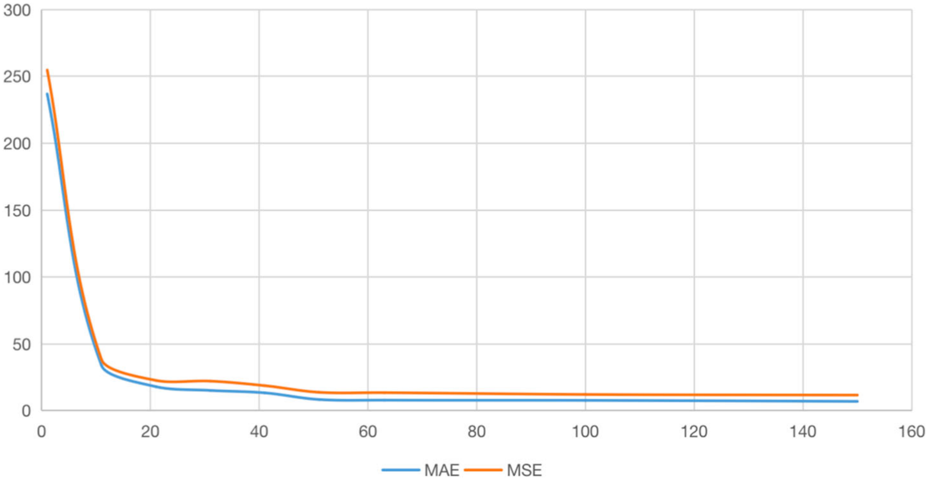

Fig. 7 The cure lines of MAE and MSE

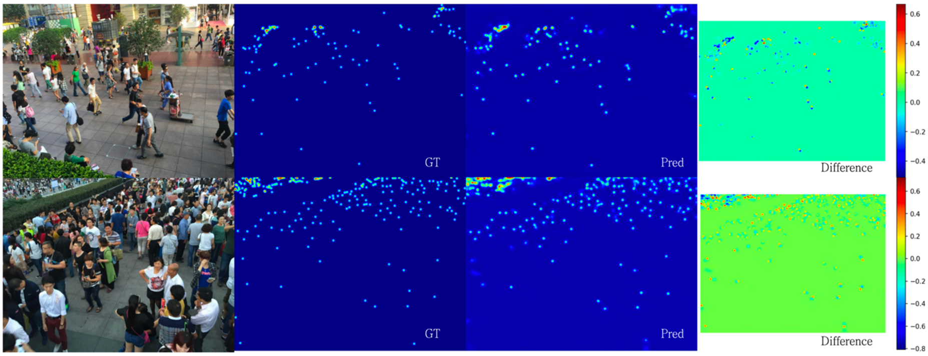

Fig. 8 From left to right, the second one is a ground truth density map, the third one is a predicted density map. The forth is the difference between predict and ground truth. “GT” denotes ground truth density map. “Pred” means predicted density map. “Difference” means difference of ground truth and predicted density map. The darker the color, the greater the difference

5 Conclusion

In this paper we propose a novel network that simultaneously focuses on building the global structure and local details of crowd distribution to generate higher quality density maps. To improve the effectiveness of mapping features to density maps, CARAFE is applied as an efficient upsampling mechanism. This work proposes the global distribution branch for generating high-quality density maps from global structures, and introduces contextual module to capture global and local features and to understand contextual information. Through the design of connecting receptive fields of different sizes, more effective contextual information can be captured. In addition, the adaptive spatial encoder module helps to cope with the distortion caused by the diverse crowd distribution and perspective. The algorithm is demonstrated on four challenging counting datasets with state-of-the-art performance. Last error label of such a large-scale scene is relatively large, and the influence of labeling noise on the model is relatively bad. The noise of label hinders the model’s ability to learn.

Acknowledgment This work was supported by the National Key Research and Development Program of China under grant of 2018******4402 and 2020YFB1712401.

Data Availability Data sharing is not applicable to this paper, as only public datasets were used during the current study and no new datasets were generated.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Babu Sam D, Surya S, Venkatesh Babu R (2017) Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5744–5752
2. Boominathan L, Kruthiventi SS, Babu RV (2016) Crowdnet: a deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM international conference on multimedia, pp 640–644

3. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750
4. Chan AB, Liang Z-SJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–7
5. Chan AB, Vasconcelos N (2009) Bayesian poisson regression for crowd counting. In: 2009 IEEE 12th international conference on computer vision, pp 545–551
6. Dai J et al (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773
7. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) vol 1, pp 886–893
8. Enzweiler M, Gavrila DM (2008) Monocular pedestrian detection: survey and experiments. IEEE transactions on pattern analysis and machine intelligence 31(12):2179–2195
9. Fu M et al (2015) Fast crowd density estimation with convolutional neural networks. Eng Appl Artif Intell 43:81–88
10. Gao J, Wang Q, Yuan Y (2019) Scar: spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
12. Hossain M, Hosseinzadeh M, Chanda O, Wang Y (2019) Crowd counting using scale-aware attention networks. In: 2019 IEEE winter conference on applications of computer vision (WACV), pp 1280–1288
13. 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, pp 2547–2554
14. Idrees H et al (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
15. Jiang X et al (2019) Crowd counting and density estimation by trellis encoder-decoder networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6133–6142
16. Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 1, 878–885
17. Li M, Zhang Z, Huang K, Tan T (2008) Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: 2008 19th international conference on pattern recognition, pp 1–4
18. Li Y, Zhang X, Chen D (2018) Csrnet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1091–1100
19. Liang D, Chen X, Xu W, Zhou Y, Bai X (2022) Transcrowd: weakly-supervised crowd counting with transformers. Science China Information Sciences 65(6):1–14
20. Liu N et al (2019) Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3225–3234
21. Liu W, Salzmann M, Fua P (2019) Context-aware crowd counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5099–5108
22. Ma Z, Wei X, Hong X, Gong Y (2019) Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6142–6151
23. Ng PC, Henikoff S (2003) Sift: predicting amino acid changes that affect protein function. Nucleic acids research 31(13):3812–3814
24. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1520–1528
25. Oh M-h, Olsen P, Ramamurthy KN (2020) Crowd counting with decomposed uncertainty. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 11799–11806
26. Onoro-Rubio D, López-Sastre RJ (Springer, 2016) Towards perspective-free object counting with deep learning. In: European conference on computer vision, pp 615–629
27. Shen Z et al (2018) Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5245–5254
28. Shu W, Wan J, Tan KC, Kwong S, Chan AB (2022) Crowd counting in the frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19618–19627
29. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations

30. Song Q et al (2021) To choose or to fuse? Scale selection for crowd counting. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 2576–2583
31. Thanasutives P, Fukui K-i, Numao M, Kijsirikul B (2021) Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In: 2020 25th international conference on pattern recognition (ICPR), pp 2382–2389
32. Topkaya IS, Erdogan H, Porikli F (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
33. Walach E, Wolf L (2016) Learning to count with cnn boosting. In: European conference on computer vision, pp 660–676
34. Wang C, Zhang H, Yang L, Liu S, Cao X (2015) Deep people counting in extremely dense crowds. In: Proceedings of the 23rd ACM international conference on multimedia, pp 1299–1302
35. Wang J et al (2019) Carafe: content-aware reassembly of features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3007–3016
36. Wang Q, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8198–8207
37. Wang Q, Gao J, Lin W, Li X (2020) Nwpu-crowd: a large-scale benchmark for crowd counting and localization. IEEE transactions on pattern analysis and machine intelligence 43(6):2141–2149
38. Xie J et al (2022) Multi-scale attention recalibration network for crowd counting. Applied Soft Computing, pp 108457
39. Xiong H et al (2019) From open set to closed set: counting objects by spatial divide-and-conquer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8362–8371
40. Xu C et al (2019) Learn to scale: generating multipolar normalized density maps for crowd counting. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8382–8390
41. Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In: 4th international conference on learning representations, pp 1–13
42. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 833–841
43. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 589–597
44. Zhang A et al (2019) Relational attention network for crowd counting. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6788–6797
45. Zhou JT et al (2021) Locality-aware crowd counting. IEEE transactions on pattern analysis and machine intelligence 44(7):3602–3613

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

Affiliations

Chengming Liu[1] · Guanzhong Hu[1] · Yinghao Li[1] · Yufei Gao[1] · Lei Shi[1]

Guanzhong Hu
hug323081@gmail.com

Yinghao Li
yinghaoli@zzu.edu.cn

Yufei Gao
yfgao@zzu.edu.cn

Lei Shi
shilei@zzu.edu.cn

[1] School of Cyber Science and Engineering, Zhengzhou University,
97 of Wenhua Street, 450002, Zhengzhou, Henan, China