Neural Computing and Applications (2020) 32:5105–5116
https://doi.org/10.1007/s00521-018-3954-7

ORIGINAL ARTICLE

Crowd density estimation based on classification activation map and patch density level

Liping Zhu[1,2] · Chengyang Li[1,2] · Zhongguo Yang[3,4] · Kun Yuan[5] · Shang Wang[1,2]

Received: 22 August 2018 / Accepted: 18 December 2018 / Published online: 3 January 2019

Abstract

The task of crowd counting and density map estimation is riddled with many challenges, such as occlusions, non-uniform density, intra-scene and inter-scene variations in scale and perspective. Due to the development of deep learning and large crowd datasets in recent years, most crowd counting methods have achieved notable success. This paper aims to solve crowd density estimation problem for both sparse and dense conditions. To this end, we make two contributions: (1) a network named Patch Scale Discriminant Regression Network (PSDR). Given an input crowd image, it divides the image into patches and sends image patches of different density levels into different regression networks to get the corresponding density maps. It combines all patch density maps to predict the entire density map as the output. (2) A person classification activation map (CAM) method. CAM provides person location information and guides the generation of the entire density map in the final stage. Experiment confirms that CAM allows PSDR to gain another round of performance boost. For instance, on the SmartCity dataset, we achieve (8.6–1.1) MAE and (11.6–1.4) MSE. Our method combining above two methods performs better than state-of-the-art methods.

Keywords Crowd density estimation · Image patch · Density level · Attention mechanism · Classification activation map

1 Introduction

The stampede is easy to happen in various public places, such as supermarkets, subways, train stations and other public places. Thus, it is of great value to carry out effective crowd density estimation and crowd aggregation detection. The automatic detection of crowd density and distribution in public places by video plays an important role in the prevention of potential security risks. The monitoring video cameras in the natural scene for specific applications are generally high and far away. This will lead great challenges to effective crowd density estimation, due to large perspective effects, large light changes, big noise and so on.

There is a certain inclination angle between the monitoring camera and the horizontal plane in the monitoring scene. The captured images have the following phenomenon due to the perspective effects: (1) The effect of "Small in the distance, large in the vicinity" is formed, as shown in Fig. 1. The distance to cameras is inversely related to person pixels occupying the image; (2) the crowd in the distance is gathered, leading to high concentration.

✉ Chengyang Li
2017215536@student.cup.edu.cn

Liping Zhu
zhuliping@cup.edu.cn

Zhongguo Yang
yangzhongguo@ncut.edu.cn

Kun Yuan
kyuan033@uottawa.ca

Shang Wang
2016215057@student.cup.edu.cn

[1] College of Information Science and Engineering, China University of Petroleum (Beijing), Beijing, China

[2] Key Lab of Petroleum Data Mining, China University of Petroleum (Beijing), Beijing, China

[3] Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, North China University of Technology, Beijing, China

[4] School of Computer Science, North China University of Technology, Beijing, China

[5] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

Fig. 1 Persons at different shooting distances

This will increase the difficulty of crowd density estimation. The existing algorithms mainly deal with the above problems based on detection and regression, but they are not satisfactory in accuracy. Detectors [1, 2] perform poorly in sparse street scenes, due to few person pixels and serious occlusion. (3) The existing mass density estimation algorithms [3–5] based on regression are not accurate in 20–50 people scenes. In some real scenes, such as elevators, streets and flyovers, crowd density must be mastered. It can provide information of crowd distribution and abnormal crowd flow in time. In summary, crowd density estimation for sparse scenes is a hot topic and difficult problem.

In this paper, we aim to conduct accurate crowd counting from an arbitrary image, with an arbitrary camera perspective and crowd density. To overcome above challenges, we propose Patch Scale Discriminant Regression Network (PSDR) with person classification activation map (CAM), as shown in Fig. 2. Contributions of this paper are summarized as follows:

1. First, we propose a network named Patch Scale Discriminant Regression Network (PSDR). PSDR takes a whole image as the input and outputs a density map whose integral gives the overall crowd count. When designing density levels, we use a density classification strategy which is closest to the real situation. The experiment shows that using image patches for scales has better performance than the whole image.
2. Second, we propose a person classification activation map (CAM) method to improve the entire density map prediction. The motivation is that the information of person heads at the image patch's edge is missing because of image patch segmentation. Therefore, we add the person CAM into our model. The person CAM makes the model focus on the human head area. Experiments prove that the person CAM can improve the performance of PSDR.

The rest of our paper is structured as follows. Section 2 presents previous works of crowd density prediction, switch-based CNN structure and classification activation map. Section 3 introduces our proposed method, while Sect. 4 presents the experimental results of different datasets. In Sect. 5, we make a conclusion of the paper.

2 Related work

2.1 Crowd density estimation

Crowd density estimation is the estimation of crowd distribution and specific person number. The density map provides crowd distribution information and statistical characteristics in the picture, which is as important as total person number. The CNN structure has achieved great success in image processing [6, 7], as is the crowd density prediction [8–14]. Current crowd density estimation methods are mainly based on detection or regression. Detection methods are applicable to scenarios which have less people and no occlusion, such as detectors based on

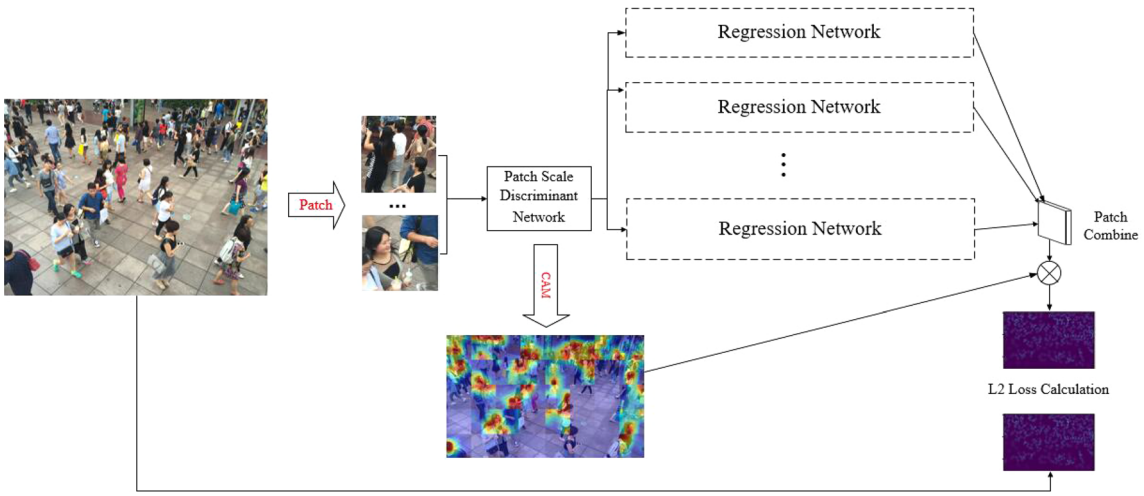

Fig. 2 The framework of Patch Scale Discriminant Regression Network and CAM

adjacent frames. Other methods based on regression can be divided into two categories. One is to detect handmade characters in the image, such as edge feature and texture feature. Then, regression function is chosen to estimate total person number. The other is based on density map regression and deep neural networks, which has become the current best method of crowd density estimation.

Multi-column CNN used by references [5, 15] fuses features from different CNN columns to regress the crowd density map. Several CNN columns [5] with different receptive fields are used to capture the large variation in scale and perspective in crowd scenes. CrowdNet [15] uses a VGG network employing dilated layers complemented by a shallow network with different receptive field and field of view. Both models fuse the feature maps from the CNN columns by entering a 1*1 convolutional layer to predict the crowd density map. However, the weighted averaging technique is global and does not take local density variation into account. Therefore, we use image patches for training to learn local information, not the whole image. Image patches provide richer local information and obtain a more accurate density map by regression.

2.2 Switch-based CNN

Switch-based CNN is first used in object recognition. To improve single-object image classification, Surva et al. [16] present SwiDeN, a CNN structure which recognizes objects regardless of how they are visually depicted. In SwiDeN, a novel deep depictive style-based switching mechanism is utilized. It appropriately addresses the depiction-specific and depiction-invariant aspects of the problem. The switch-based design not only reduces the overall burden of the generalized object recognition task, but also enables the system to address depiction-specific and depiction-invariant aspects of the problem.

Similarly, this idea can be extended to density map prediction. Switch-CNN [3] consists of three CNN regressors with different architectures and a switching classifier. It selects the optimal regressor for an input crowd scene patch. Though local information can be obtained by patches, the connection between each image patch is weakened. Also, image patches in Switch-CNN are simply divided into three categories, without explaining the specific classification method and basis. Bad image patch classification may lead to learn redundant information by networks. Therefore, we propose a more sophisticated Patch Scale Discriminant Network which divides patches into six levels according to person number in patches. All density maps obtained from regression networks are merged into the whole predicted density map. It finally combines the global person classification activation map to make up for the missing correlation between image patches.

2.3 Classification activation map

In computer vision, attention mechanism is applied to a variety of problems, including image classification, segmentation, action recognition, image captioning and visual question answering. For example, in the context of medical image analysis, attention models have been exploited for medical report generation as well as joint image and text classification.

Zeiler & Fergus presents what a CNN learns in [17]. However, their method only involves significant computations to generate this understanding. Zhou et al. [18] showed that various layers of CNN behave as unsupervised object detectors by a new technique called CAM (class activation mapping). By using global average pooling [19] layer and visualizing the weighted combination of the resulting feature maps at the penultimate (pre-softmax) layer, they were able to obtain heatmaps that explain which parts of an input image were looked at by CNN for assigning a label. Yet, this technique involves retraining a linear classifier for each class. Selvaraju et al. [20] came up with an efficient generalization of CAM, known as Grad-CAM, which fuses the class discriminative property of CAM with existing pixel-space gradient visualization techniques such as Guided Backpropagation [21] and Deconvolution [17] to highlight fine-grained details on the image. Therefore, Grad-CAM makes CNN-based models more transparent by visualizing input regions with high-resolution details that are important for predictions.

CAM based on attention mechanisms allows the model to focus on the most relevant features and locations as needed. Crowd density estimation is similar to semantic segmentation [22]. Therefore, in view of the outstanding performance of attention models in semantic segmentation, we introduce person CAM in the proposed method. It can improve the performance of density maps by providing person characteristics and location.

3 Proposed method

Using crowd density estimation algorithm for reference, we design a crowd density estimation algorithm based on image patch and regression. The proposed method regresses the image to the crowd density distribution and then estimates the person number in the density map. The crowd density distribution map is learned from several labeled crowd images, and the person number of crowds is obtained through summing all pixel values.

3.1 Crowd density map ground truth

Person pixels corresponding to different samples have different sizes in 3D scenes, due to perspective and distortion. To accurately estimate crowd density, we need to consider the distortion caused by the angle between the ground plane and the image plane. This perspective angle is difficult to obtain. Therefore, the average distance to k-nearest neighbors can be used as an assessment of geometric distortion. Specifically, for each person's head x_i, the distance between the head and other k-nearest neighbors is calculated to be $\{d_1^i, d_2^i, \ldots, d_k^i\}$; then, the mean distance $\bar{d}_i = \frac{1}{k}\sum_{j=1}^{k} d_j^i$ is obtained. Final crowd density map $F(x)$ is produced, as shown in formula 1.

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), \quad \sigma_i = \beta \bar{d}_i \tag{1}$$

Here, $\delta(x - x_i)$ is the delta function, $G_{\sigma_i}(x)$ is the Gauss function whose variance is σ_i, and the empirical value β is set to be 0.3. The density map ground truth is generated based on the whole graph, using the empirical parameters. One sample is shown in Fig. 3.

3.2 Image patch

In the direction of the camera, pixel number of human body or head occupying the image is different, as shown in Fig. 4. The camera angle is difficult to obtain, so it is difficult to calculate perspective. The method using perspective cannot solve the problem of human body size disagreement. The whole image has a large perspective angle, but for the local parts of an image, the corresponding perspective can be determined by the person number.

In the theory of image classification, the object itself constitutes a hierarchical structure in the image. It is difficult to find all information on one scale. Therefore, image patches are hierarchically organized. Accuracy can be effectively improved by learning information at image patches of each scale. Small patches can capture more image details, but lose the details of patches. Large patches can describe a wider range of image details.

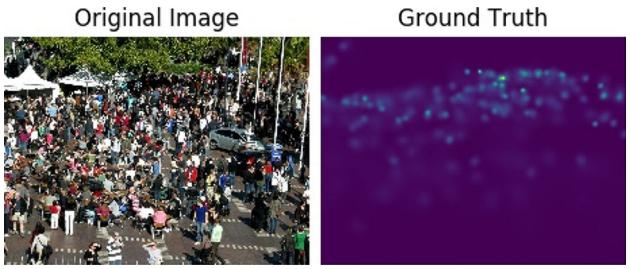

Fig. 3 The original image and generated ground truth

Fig. 4 Person heads under different shooting angle

Considering all above, the whole image is cut into several patches of 128*128 pixels. Each patch can be considered at the same perspective level. There is a certain overlap area between image patches, in order not to cause missing of human head features at the edge position. Thus, it can reduce the influence caused by the perspective of the camera. It also ensures that there is no obvious difference in size of each person in each image patch. Image patches are shown in Fig. 5.

Many studies take the scale of crowd density into account, that is, different crowd density should be based on different regression models. Each image patch corresponds to different crowd distribution. The better the image patch is divided, the more easily the relationship between person pixels and person number in each image patch performs. A more refined density estimation structure can be obtained by training regression models separately for different crowd levels.

Fig. 5 Image patches of different density levels

3.3 Patch Scale Discriminant Network

After obtaining image patches, the crowd density is divided into six levels [23]: extremely sparse, very sparse, sparse, medium, dense and very dense. Different levels have different person number per image, as shown in Table 1.

In many studies [3, 24, 25], they take the scale of crowd density as an important consideration. The main approach is to use convolution kernels of different scales. This can achieve the crowd density response of different scales. Local patch level in our proposed method not only effectively overcomes the perspective problem of cameras, but also reduces the dependence on training samples. Patch Scale Discriminant Network is trained, whose input is image patches of 128*128 pixels. Its output is six different levels, respectively, corresponding to crowd density distribution under different scenes. The higher crowd density distribution is, the fewer pixels a single human body occupies. Patch Scale Discriminant Network is based on a VGG16 model. We ensure that the size of human body in each image patch is maintained on the same scale. Different image patches can obtain more accurate and more detailed density maps. The prediction results are shown in Fig. 6.

Image patches at bottom of the image have the lowest level, as shown in Fig. 6. The higher image patch is, the higher density level is. The perspective angle causes the crowd population to be very different. Similarly, the distribution of the corresponding density maps is also different. Therefore, it is feasible to deal with the image patches of each density level separately.

3.4 Global person CAM

Convolutional features naturally retain spatial information which is lost in fully connected layers, so the last convolutional layers can be expected to have the best compromise between high-level semantics and detailed spatial information. The neurons in these layers look for semantic class-specific information in the image. A CAM sensitive to person characteristics stands for crowd density heatmap to some extent. Referring to the research [18], person CAM can be used as the heatmap of person distribution. Grad-

Fig. 6 Density level predictions of each image patch in sparse and dense scene

CAM [20] uses the gradient information flowing into the last convolutional layer of the CNN to understand the importance of each neuron for a decision of the interest. Therefore, we use this method to obtain person localization heatmap.

Firstly, the gradients y^{person} of the score for class person are first computed. Then, the gradients of A^k are computed through back propagation, i.e., $\frac{\partial y^{\text{person}}}{\partial A^k}$.

Here, A^k represents feature maps of a k-channel convolutional layer. Next, these gradients are global average pooled to obtain the neuron importance weights α_k^{person}, as shown in formula 2. This weight α_k^{person} represents a partial linearization of the deep network downstream from A and captures the importance of feature map k for class person. Finally, person CAM is obtained by combining activation maps with weights.

$$CAM^{\text{person}} = \sum_k \alpha_k^{\text{person}} A^k \tag{2}$$

In this paper, we first train Patch Scale Discriminant Network (Sect. 3.3) to convergence, which can divide image patches into six different levels. This model includes two stages: feature extraction and classification. Then, based on the extracted feature by network, the heatmap is obtained by visualizing the last convolutional layer. Some of the results are shown in Fig. 7. It indicates that CAM outputs a weighted map which weights features pixel by pixel. In both sparse and dense scenes, most of the human body can be highlighted in the image, regardless of human body's size. In addition, the boundaries of highlighting subregions are distinct, which fully indicates that the global features of human body can be introduced by

Table 1 Density levels of image patches

Crowd density (person/pixel)	Density level	Density level (person number)	Level number
0	Extremely sparse	0	0
2/128*128	Very sparse	(1, 2]	1
5/128*128	Sparse	(2, 5]	2
10/128*128	Medium	(5, 10]	3
20/128*128	Dense	(10, 30]	4
60/128*128	Very dense	(30, $+\infty$)	5

Fig. 7 The person CAM, even effective for small and seriously obscured persons

CAM. The knowledge learned from Patch Scale Discriminant Network model can be perfectly migrated to the person response area.

3.5 Crowd density regression

The image patches divided by a single picture get the corresponding level after the trained Patch Scale Discriminant Network. Then, they enter the corresponding regression network according to the predicted level to predict the density map. When the corresponding density map of all patches is obtained, the whole density map is obtained by splicing all the density maps. Every regression network should be pre-trained early.

Since MCNN [5] has great performance in density prediction, we take it as a reference. Besides, crowd density map makes a prediction at every pixel, whether it is a person or not. It can be considered as a pixel-wise task, just like semantic segmentation. FCN [26] can efficiently learn to make dense predictions. It also shows that pooling layers lose pixel information, leading to a bad effect on pixel-wise segmentation. Therefore, we remove the pooling layers so that the structure becomes a fully convolutional network, as shown in Fig. 8.

Above all, we design a crowd density estimation framework by combining Patch Scale Discriminant Network and global person CAM, as shown in Fig. 2. The density map ground truth is accumulated by the Gauss distribution based on the head position and the distribution distance of the surrounding human head. In the process of loss calculation, the L2 loss of predicted density map and original density map is adopted, as shown in formula 3.

$$L_D(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{p=1}^{P}||F(X_i(p);\Theta) - F_i(p)||_2^2 \tag{3}$$

Here, N is the number of batch training samples, and $F_i(p)$ is the density value of the pixel point p in the annotation picture i. There is a certain connection between the crowd density map and the specific number of the crowd. Therefore, the density map can be converted into the person number, directly using the integral method.

4 Experiment

4.1 Evaluation metric

The evaluation indexes of crowd density estimation include MAE (Mean Absolute Error) and MSE (Mean Squared Error). MAE calculation is shown in formula 4. MSE calculation is shown in formula 5.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y - y'| \tag{4}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt[2]{|y - y'|^2} \tag{5}$$

Here, y is the actual person number, while y' is the predicted number of people in the experiment. Note that y' is the sum of the values of all the pixels on the predicted density map. Roughly speaking, MAE indicates the accuracy of the estimation, and MSE indicates the robustness of the estimation.

We also use the PSNR [27] (Peak Signal-to-Noise Ratio) and SSIM [28] (Structural Similarity in Image) to evaluate the quality of the output density map. To calculate the PSNR and SSIM, we follow the preprocess given by [29], which includes the density map resizing (same size

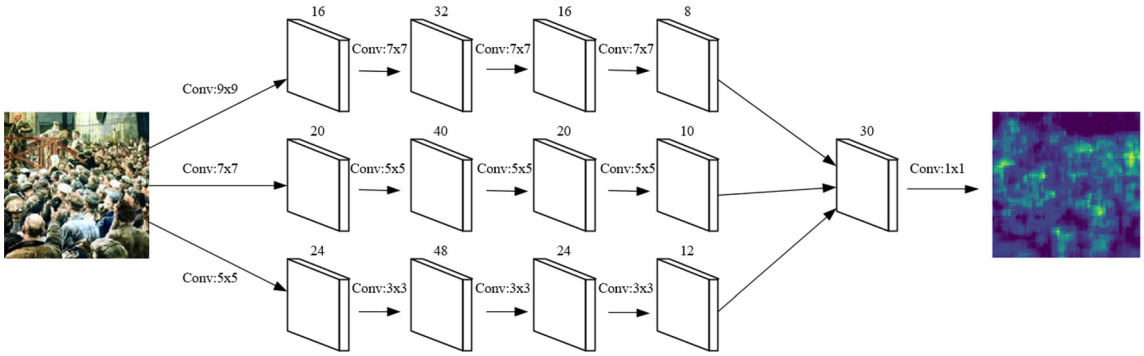

Fig. 8 The structure of density regression network

with the original input) with interpolation and normalization for both ground truth and predicted density map.

4.2 Dataset

ShanghaiTech dataset The ShanghaiTech dataset [5] is composed of 1198 images with a total amount of 330,165 persons. The dataset is divided into two parts: PartA and PartB. Images in PartA are highly congested scenes randomly downloaded from the Internet, while images in PartB are relatively sparse crowd scenes taken from streets. In PartA, 300 images are used for training, and 182 images are used for testing. In PartB, 400 images are used for training and 316 images are used for testing. Some images are shown in Fig. 9.

SmartCity dataset The dataset [4] contains 50 images, mostly collected in urban scenes, such as office entrance, the sidewalk, atrium, shopping center and so on. These images are captured by high angle video surveillance equipment. There are only a few people in SmartCity dataset. The crowd in the images is very sparse. The average number of pedestrians in an image is only 7.4, the minimum value is 1, and the maximum is 14. Therefore, the SmartCity dataset can be used to test the generalization ability of the algorithm under the unusual sparse crowd scene. Some images are shown in Fig. 10.

UCF_CC_50 dataset UCF_CC_50 dataset [30] includes 50 images with different perspective and resolutions. The number of annotated persons per image ranges from 94 to 4543 with an average of 1280. Fivefold cross-validation is performed following the standard setting in [30]. Some images are shown in Fig. 11.

The UCSD dataset The UCSD dataset [31] has 2000 frames captured by surveillance cameras. These scenes contain sparse crowd varying from 11 to 46 persons per image. The region of interest (ROI) is also provided. The resolution of each frame is fixed and small (238*158). Among the 2000 frames, we use frames 601 through 1400 as training set and the rest of them as testing set according to [31]. Some images are shown in Fig. 12.

In summary, the details of datasets, including size of dataset, resolution ration, and minimum, maximum and average people number, as shown in Table 2. It shows that

Fig. 9 Samples of ShanghaiTech dataset (PartA and PartB)

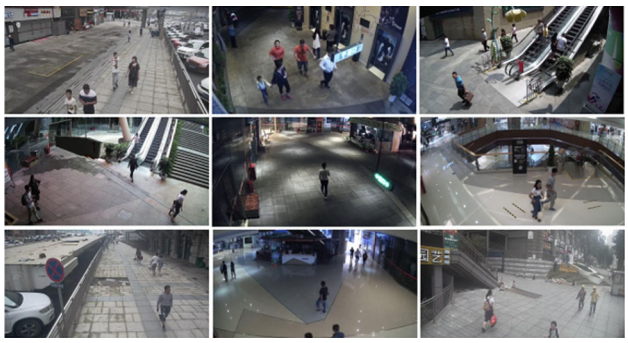

Fig. 10 Samples of SmartCity dataset

Fig. 11 Samples of UCF_CC_50 dataset

Fig. 12 Samples of The UCSD dataset

SmartCity and UCSD datasets are especially crowd density estimation for sparse scenes.

4.3 Training process

All images are set to 1024*768, so that they can be cut into 48 patches of 128*128. Similarly, the corresponding density map is operated in the same way. Training images are divided into patches randomly. After data preprocessing, we design a training procedure, as illustrated in Algorithm 1.

1. First, the density level of each image patch is obtained by the labeled head coordinates. According to image patches and the corresponding density level, Patch

Table 2 Statistics information of datasets

Dataset	Resolution ratio	Image number	Minimum people number	Maximum people number	Average people number
ShanghaiTech					
PartA	Different	482	33	3139	501.4
PartB	768*1024	716	9	578	123.6
SmartCity	1920*1080	50	1	14	7.4
UCF_CC_50	Different	50	94	4543	1279.5
UCSD	238*158	2000	11	46	25

Scale Discriminant Network (PSD) is trained to convergence.

2. Second, the person CAM of all images is obtained by the trained PSD model, using Grad-CAM method.
3. Third, each density level corresponds to an independent density map regression network. Therefore, the regression networks are trained by image patches of each level separately.
4. Finally, the parameters of pre-trained PSD and regression models of each level are loaded into the proposed whole network. All the layers are allowed to be updated. The whole network is fine-tuned for overall optimization.

Algorithm 1: Training process for PSDR + CAM

Step 1. Patch Scale Discriminant Network is trained to convergence with six levels.

Step 2. Person CAM of all training samples is obtained by the trained PSD model.

Step 3. Six regression networks for six levels are trained until convergence.

Step 4. The whole network is fine-tuned to convergence for overall optimization.

We use one NVIDIA Tesla K80 GPU and PyTorch as the platform. Training samples are from ShanghaiTech PartA. In Step 1, PSD model is pre-trained on ImageNet [32]. In Step 3, for each regression network, we set batchsize to 64. The learning rate is initialized at 1e−3 and decayed to 1e−5 when loss tends to be constant. Step 4 converges very fast, requiring about 20 epochs in total. The learning rate of whole model is set to 1e−5.

4.4 Performance evaluation

In order to verify the effectiveness of our proposed method, we test it on five datasets introduced in Sect. 4.2. We compare our methods with state-of-the-art methods for crowd counting.

Using image patches performs better than whole image Image patches follow the idea of divide-and-conquer in data structure. First, the original problem of density map prediction of the whole image is decomposed into several subproblems of the density map predictions of image patches. These subproblems are small examples of the original problem. Then, these subproblems are solved by using regression networks. Finally, the solutions of these subproblems (the predicted density maps of each image patch) are merged into the solution of the original problem (the predicted density map of the whole image).

Original image patches and predicted density maps are shown in Fig. 13. It can be observed that in lower level image patches, model can learn the contour information of the person head; in the image patches of higher level, the model mostly learns the distribution of person heads. By combining the knowledge learned from different levels of models, we can better handle the crowd counting problem in sparse cases. Unlike previous regression models, which equate sparse crowd (large heads) and dense crowd (small heads), our method takes these two different situations into account and obtains more detailed density maps.

Good results are achieved in the patches of ShanghaiTech Part A/B. The density map of each level patch is regressed, and the crowd number is obtained based on density map. The prediction results of some images are shown in Fig. 14. This indicates that the method makes full use of the regression model of local image patches and accurately extracts the more detailed human features. At the same time, after combining the global person CAM, the missing of the edges between image patches is compensated. Also, the locations of person in CAM help regression network to learn in a more precise direction.

Comparison with state-of-the-art methods The comparisons of different methods on the five datasets (listed in Table 2) are shown in Tables 3, 4 and 5. Our method reduces the MAE error from 8.6 to 7.5 and the MSE error from 11.6 to 10.2 on SmartCity dataset. Besides, our

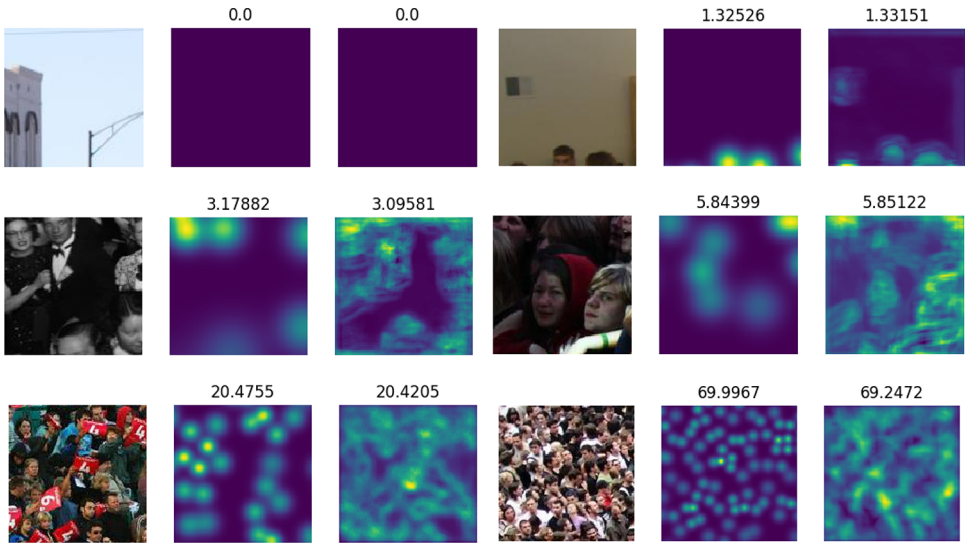

Fig. 13 Original image patches, the corresponding ground truth and density predicted results (from left to right)

method achieves (1.04–0.01) MAE on UCSD dataset. However, our method on dense scenes (ShanghaiTech PartA and UCF_CC_50) works worse than state-of-the-art methods, such as ACSCP [33], M-task [34] and D-CNet [35]. It means our method works better in sparse scenes but a bit worse in dense scenes. On ShanghaiTech PartB dataset, the performance of our method is very close to the best method M-task [34]. From the perspective of dataset population density distribution, the sparser the crowd is, the better our method performs. We find that this happens due to classification activation map (CAM).

The principle of CAM is to visualize the area that the model notices when it is judged as a category. In sparse scenes, CAM will give the location and information of each person's head. However, CAM may miss some person information in dense scenes. The comparison is shown in Fig. 7. When the density level is high, the PSD network (Sect. 3.3) can make the classification judgment when it takes attention to the number of heads required for classification. This may cause the loss of some people's information in the CAM. For example, when discriminating as the fifth category (greater than 30 people), the model can make a classification judgment, only finding 30 people.

Table 3 Density estimation results of different methods on ShanghaiTech dataset

Method	PartA		PartB	
	MAE	MSE	MAE	MSE
MCNN [5]	110.2	1173.2	26.4	41.3
SwitchNet [3]	90.4	135.0	21.6	33.4
SaCNN [4]	86.8	139.2	16.2	25.8
CP-CNN [29]	73.6	106.4	20.1	30.1
ACSCP [33]	75.7	**102.7**	17.2	27.4
M-task [34]	73.6	112.0	**13.7**	**21.4**
D-CNet [35]	**73.5**	112.3	18.7	26.0
Our method	84.2	128.6	14.3	23.9

Bold values indicate the highest level at present

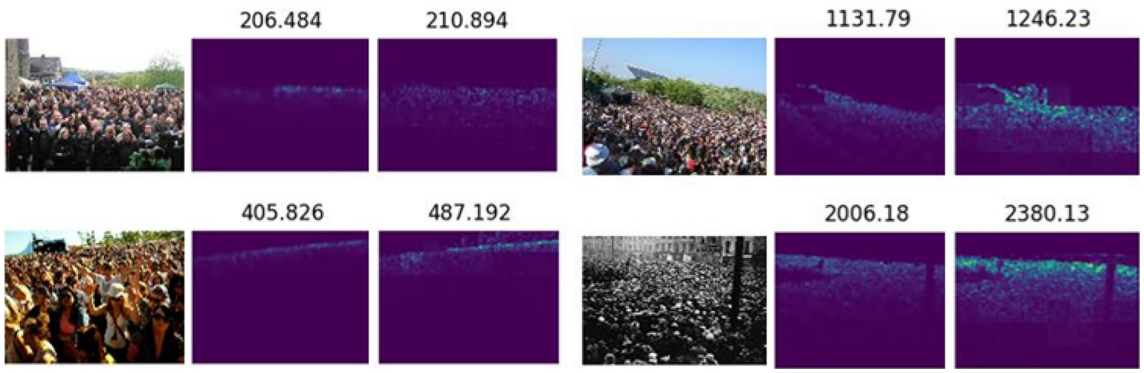

Fig. 14 The original image, ground truth, predicted density map (from left to right)

Table 4 Density estimation results of different methods on SmartCity dataset

Method	SmartCity	
	MAE	MSE
MCNN [5]	40.0	46.2
SwitchNet [3]	23.4	25.2
SaCNN [4]	8.6	11.6
Our method	**7.5**	**10.2**

Bold values indicate the highest level at present

Table 5 Density estimation results of different methods on UCF_CC_50, UCSD datasets

Method	UCF_CC_50		UCSD	
	MAE	MSE	MAE	MSE
MCNN [5]	377.6	509.1	1.07	**1.35**
SwitchNet [3]	318.1	439.2	1.62	2.10
SaCNN [4]	314.9	424.8	–	–
CP-CNN [29]	295.8	**320.9**	–	–
ACSCP [33]	291.0	404.6	1.04	**1.35**
M-task [34]	**279.6**	388.9	–	–
D-CNet [35]	288.4	404.7	–	–
Our method	302.3	411.6	**1.03**	1.37

Bold values indicate the highest level at present

4.5 Ablation Study on ShanghaiTech PartA

In this subsection, we perform an ablation study to demonstrate the effects of different modules in the proposed method. Each module is added sequentially to the network, and results for each configuration are compared on ShanghaiTech PartA dataset. Due to the presence of large variations in density, scale and appearance of people across images in this dataset, estimating the count with high degree of accuracy is difficult. Thus, this dataset was chosen for the detailed analysis of performance of the proposed architecture.

Table 6 Estimation errors and density map quality for different configurations on ShanghaiTech PartA dataset

Method	MAE	MSE	PSNR	SSIM
MCNN [5]	110.2	173.2	21.4	0.52
Regression network	107.7	180.5	21.1	0.49
PSDR	85.9	132.4	21.5	0.61
PSDR + CAM	84.2	128.6	21.59	0.64

Following three configurations are evaluated: (1) Regression Network: Only use the regression network (Fig. 8) in Sect. 3.5. (2) PSDR: Regression network with Patch Scale Discriminant Network in Sect. 3.3. (3) PSDR + CAM: Use classification activation map to guide the output of PSDR. This is our total structure. MAE, MSE, PSNR and SSIM are calculated and compared with MCNN.

Patch Scale Discriminant Regression Network weakens the perspective effects Without Patch Scale Discriminant Network, only regression network has high error and predicts low quality density maps. As shown in Table 6, PSDR has a huge improvement over it. MAE and MSE achieve 18.2 and 41.9 lower, respectively. This proves the effectiveness of Patch Scale Discriminant Network.

CAM improves the quality of Patch Scale Discriminant Regression Network While Patch Scale Discriminant Regression Network already has low MAE and MSE, CAM brings further improvement to it. As shown in Table 6, the improvement in MAE and MSE is − 1.7 and − 3.8,

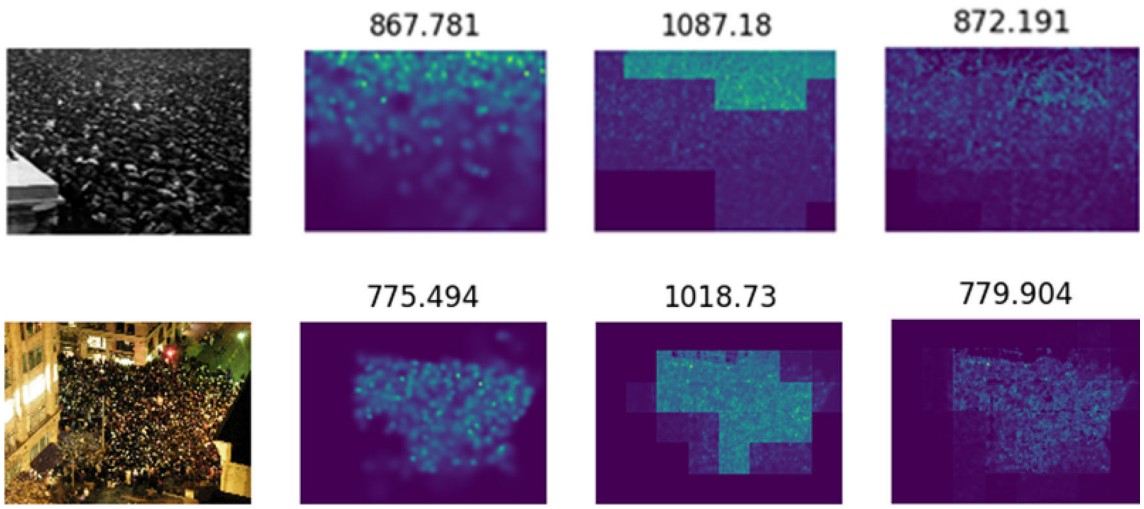

Fig. 15 Original image, density map ground truth, image predictions of PSDR and PSDR + CAM (from left to right)

respectively. Predictions of density map using model which is added CAM or not are shown in Fig. 15. It shows that the proposed CAM not only improves performance over the original model, but also allows us to diagnostically visualize the importance of features at different positions.

5 Conclusion

This paper makes two contributions to tackling crowd density map prediction problem. First, we propose Patch Scale Discriminant Regression Network (PSDR) for learning local information for six density levels. PSDR employs a patch scale strategy. Then, it learns the density map distribution information of six density levels with six regression networks, respectively. Despite the fact that PSDR is accurate and effective, it is yet to be improved. We introduce the person classification activation map (CAM) into the density map generation process of the entire image. CAM provides person location information as global information and improves PSDR considerably.

In the future, in some areas of image research, using classification activation map as a traditional feature is an important research direction.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61672042), Models and Methodology of Data Services Facilitating Dynamic Correlation of Big Stream Data, 2017.1 ~ 2020.12.

References

1. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: IEEE conference on computer vision and pattern recognition. IEEE Computer Society, pp 779–788
2. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision – ECCV 2016. European conference on computer vision, ECCV 2016. Lecture notes in computer science, vol 9905. Springer, Cham, pp 21–37
3. Sam DB, Surya S, Babu RV (2017) Switching convolutional neural network for crowd counting
4. Zhang L, Shi M, Chen Q (2018) Crowd counting via scale-adaptive convolutional neural network. In: Proceedings of the 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1113–1121
5. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016). Single-image crowd counting via multi-column convolutional neural network. In: IEEE conference on computer vision and pattern recognition. IEEE Computer Society, pp 589–597
6. Zhang H, Cao X, Ho JKL, Chow TWS (2017) Object-level video advertising: an optimization framework. IEEE Trans Ind Inf 13(2):520–531
7. Zhang H, Ji Y, Huang W, Liu L (2018) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3579-x
8. Nagao K, Yanagisawa D, Nishinari K (2018) Estimation of crowd density applying wavelet transform and machine learning. Physica A Stat Mech Appl 510:145–163
9. Zhou B, Song B, Hassan MM, Alamri A (2018) Multilinear rank support tensor machine for crowd density estimation. Eng Appl Artif Intell 72:382–392
10. 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
11. Sindagi VA, Patel VM (2018) A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognit Lett 107:3–16
12. Saleh SAM, Suandi SA, Ibrahim H (2015) Recent survey on crowd density estimation and counting for visual surveillance. Eng Appl Artif Intell 41:103–114
13. Chen K, Kämäräinen JK (2016) Pedestrian density analysis in public scenes with spatiotemporal tensor features. IEEE Trans Intell Transp Syst 17(7):1968–1977
14. Zhang C, Kang K, Li H, Wang X, Xie R, Yang X (2016) Data-driven crowd understanding: a baseline for a large-scale crowd dataset. IEEE Trans Multimedia 18(6):1048–1061
15. Boominathan L, Kruthiventi SSS, Babu RV (2016) CrowdNet: a deep convolutional network for dense crowd counting. In: ACM on multimedia conference. ACM, pp 640–644
16. Sarvadevabhatla RK, Surya S, Kruthiventi SSS et al (2016) SwiDeN: convolutional neural networks for depiction invariant object recognition. In: ACM on multimedia conference. ACM, pp 187–191
17. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision, vol 8689. Springer, Cham, pp 818–833
18. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
19. Lin M, Chen Q, Yan S (2013) Network in network. Comput Sci
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE international conference on computer vision (ICCV), 22–29 Oct 2017. IEEE, Venice, Italy, pp 618–626
21. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for simplicity: the all convolutional net. arXiv preprint, arXiv:1412.6806
22. Chen LC, Yang Y, Wang J, Xu W, Yuille AL (2016). Attention to scale: scale-aware semantic image segmentation. In: Computer vision and pattern recognition. IEEE, pp 3640–3649
23. Polus A, Schofer JL, Ushpiz A (2016) Pedestrian flow and level of service. J Transp Eng 109(1):46–56
24. Zeng L, Xu X, Cai B, Qiu S, Zhang T (2017) Multi-scale convolutional neural networks for crowd counting. In: Proceedings of the 2017 IEEE international conference on image processing (ICIP). IEEE, pp 465–469
25. Li Y, Zhang X, Chen D (2018) CSRNet: dilated convolutional neural networks for understanding the highly congested scenes
26. Shelhamer E, Long J, Darrell T (2014) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):1
27. Horé A, Ziou D (2013) Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure? IET Image Process 7(1):12–24
28. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612

29. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid cnns. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, pp 1879–1888
30. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. In: Computer vision and pattern recognition, vol 9. IEEE, pp 2547–2554
31. Chan AB, Liang ZSJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–7
32. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115(3):211–252
33. Shen Z, Xu Y, Ni B, Wang M, Hu J, Yang X (2018) Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5245–5254
34. Liu X, van de Weijer J, Bagdanov AD (2018) Leveraging unlabeled data for crowd counting by learning to rank. arXiv preprint arXiv:1803.03095
35. Shi Z, Zhang L, Liu Y, Cao X, Ye Y, Cheng MM, Zheng G (2018) Crowd counting with deep negative correlation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5382–5390