IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 1, 2019, accepted July 1, 2019, date of publication July 4, 2019, date of current version July 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926881

Crowd Counting Method Based on Convolutional Neural Network With Global Density Feature

ZHI LIU[1], YUE CHEN[1], BO CHEN[1], LINAN ZHU[1], DU WU[2], AND GUOJIANG SHEN[1]
[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[2]College of Computer Science and Technology, Jilin University, Changchun 130012, China

Corresponding author: Guojiang Shen (gjshen1975@zjut.edu.cn)

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY16F020033 and Grant LY16F020035, and in part by the National Natural Science Foundation of China under Grant 61701443.

ABSTRACT Crowd counting is an important research topic in the field of computer vision. The multi-column convolution neural network (MCNN) has been used in this field and achieved competitive performance. However, when the crowd distribution is uneven, the accuracy of crowd counting based on the MCNN still needs to be improved. In order to adapt to uneven crowd distributions, crowd global density feature is taken into account in this paper. The global density features are extracted and added to the MCNN through the cascaded learning method. Because some detailed features during the down-sampling process will be lost in the MCNN and it will affect the accuracy of the density map, an improved MCNN structure is proposed. In this paper, the max pooling is replaced by max-ave pooling to keep more detailed features and the deconvolutional layers are added to restore the lost details in the down-sampling process. The experimental results in the UCF_CC_50 dataset and the ShanghaiTech dataset show that the proposed method has higher accuracy and stability.

INDEX TERMS Global density feature, deep learning, convolutional neural network, crowd counting.

I. INTRODUCTION

Crowd counting is used to calculate the total number of people in images or video frames. The crowd counting methods can be divided into three categories: the direct counting method based on target detection, the indirect method based on feature regression and crowd counting based on deep learning. In the relevant researches based on target detection [1]–[5], Lin *et al.* [1] proposed to use Haar wavelet transform to extract the feature area of the head-like contour and build the SVM classifier to classify the feature area. Gardzinski *et al.* [2] proposed to use shape contour of body to achieve crowd detection and crowd density estimation. All of these methods are suitable for the scenes with low density crowd, but the detection accuracy will decrease in the case of high density crowd. In the relevant researches based on feature regression [6]–[10], the regression relationships between image features and the number of people are established for crowd counting. Chan and Vasconcelos [7] proposed to use low-level features and Bayesian regression to improve the robustness and adaptability of the regression model. Idrees *et al.* [8] proposed to use multiple-sources information to estimate the number of people in a single image, and the UCF_CC_50 dataset was introduced in this work.

Recently, with the rapid development of deep learning and big data [11]–[14], crowd counting methods based on deep learning are proposed gradually. Zhang *et al.* [15] proposed a cross-scene crowd counting model. It was trained alternately through two learning objectives, density map and global number. This algorithm is implemented based on single-column CNN. However, it is not suitable for the change in the scale of crowd. Zhang *et al.* [16] proposed to use the MCNN with three branch networks for crowd counting. Different receptive fields were used in each branch network, and this improved MCNN could adapt to the change in the scale of the crowd. They also introduced a new dataset ShanghaiTech for crowd counting. Boominathan *et al.* [17] proposed to combine the features of shallow and deep convolutional neural networks to improve spatial resolution. Sindagi and Patel [18] proposed a multi-task network which combined the high-level prior with the density estimation. Sam *et al.* [19] proposed Switch-CNN for crowd counting. In this network, a classifier was trained and an appropriate regressor was selected for input patches. Shi *et al.* [20] proposed to aggregate multiscale features into

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

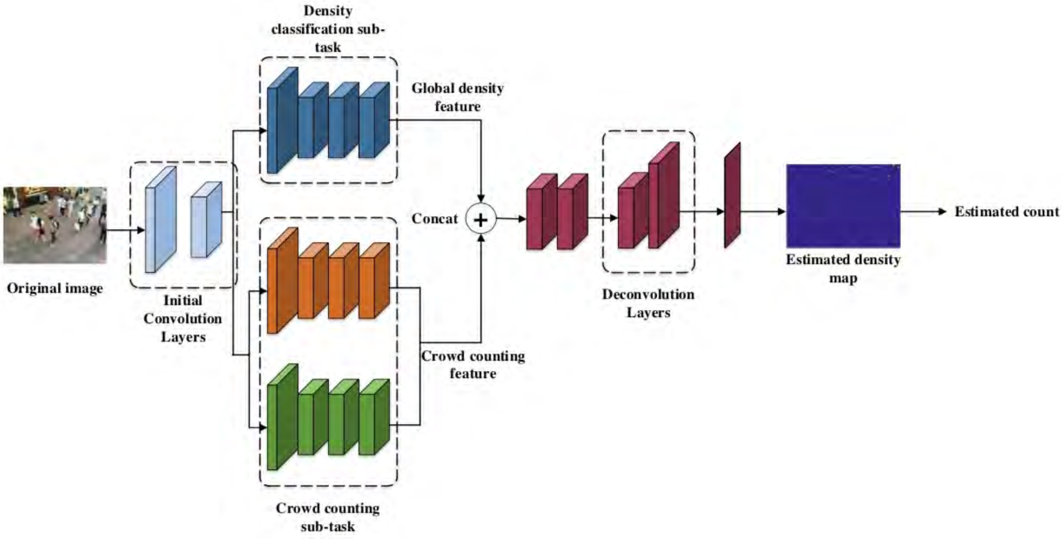


FIGURE 1. Framework of the algorithm.

a compact single vector and used deep supervised strategy to provide additional supervision signal. Fu *et al.* [21] proposed to use the LSTM structure to extract the contextual information of crowd region. Liu *et al.* [22] proposed to add an attention module to adaptively select the counting mode used for different positions on the image. Yang *et al.* [23] proposed to use the MMCNN for robust crowd counting. In this work, the location, detailed information and scale variation were taken into account to generate density map in order to improve the robustness of crowd counting method. Generally, these algorithms have good performances in the crowd counting, but the performances of these methods were not effective when the crowd distribution is uneven [24], [25].

In order to solve the problem of inaccurate counting caused by uneven crowd distribution, the global density feature is extracted and used in this paper. A convolutional neural network with global density feature by using multi-task network cascades (MNCs) [18], [26] is proposed. In order to generate a more comprehensive density map, the max-ave pooling layers are used to keep more features of the image. Meantime, the deconvolutional layers are added to the convolutional neural network in order to restore the lost details in down-sampling process. It will help to improve the accuracy of density map and further improve the accuracy of crowd counting.

II. ALGORITHM FRAMEWORK

In this paper, the main framework of the proposed method can be divided into three parts: firstly, the global density features are obtained by density classification sub-task. They are concatenated with features obtained by crowd counting sub-task. Then, the max-ave pooling is introduced to keep more features. At the same time, the deconvolutional layers are used to restore the lost details in down-sampling process. Finally, the estimated density map is generated based on feature map with global density feature. The estimated density map is used to obtain estimated counting. The framework of the algorithm is shown in Fig.1.

III. CONVOLUTIONAL NEURAL NETWORK WITH GLOBAL DENSITY FEATURE

The crowd counting method based on MCNN has achieved good counting effects so far. However, uneven crowd distributions have not been taken into account in the existed crowd counting methods. In order to solve the problem of uneven crowd distribution, the global density feature is taken into account. In this paper, the network is constructed with the two aspects: extracting feature maps with global density feature and generating a more comprehensive density map. The entire network model is shown in Fig.2.

A. FEATURE MAPS WITH GLOBAL DENSITY FEATURE

In the field of saliency detection [27], [28] and scene parsing [29], the method of combining with global features has been demonstrated effectively. It can improve the accuracy of the features. In the field of crowd density estimation [30], crowd density is quantified into five levels, and different density levels are used to describe the crowd density features. Based on these methods, this paper uses the density level as the global feature of the image. Moreover, the obtained feature map will be more accurate by combining the global density feature with the crowd counting feature.

The proposed network is mainly based on MCNN. In order to add the global density feature, the multi-task network cascades (MNCs) structure used in CMTL is applied. It is used to cascade the density classification sub-task and the crowd counting sub-task.

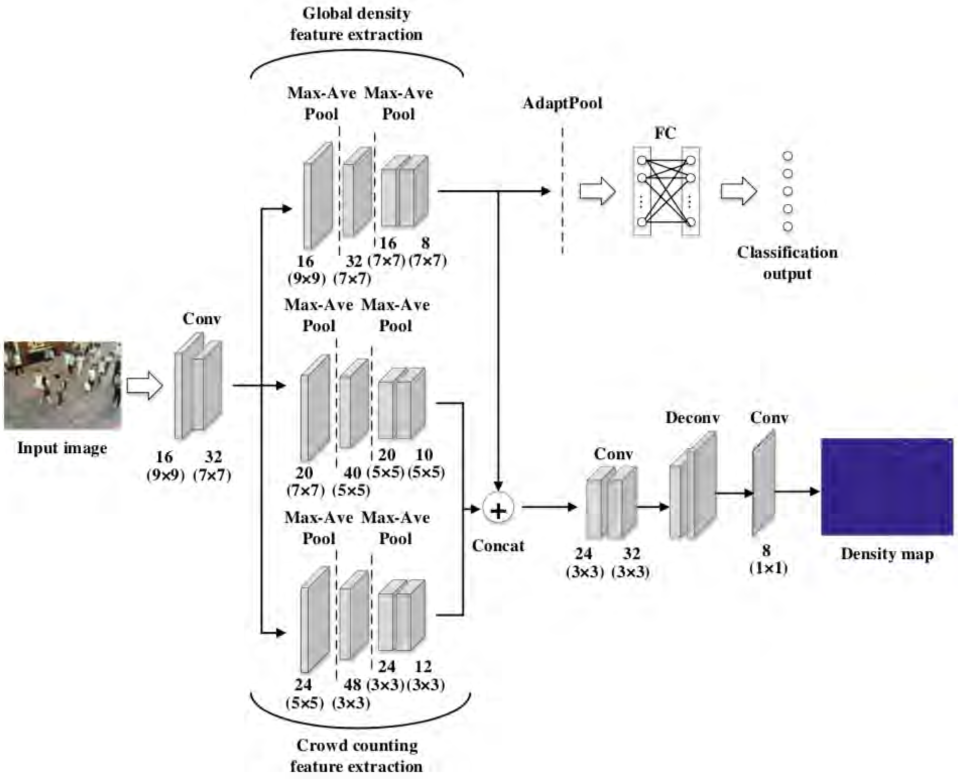


FIGURE 2. The architecture of convolutional neural network with global density feature.

The density classification sub-task is used to classify the density level. Through estimating the crowd density level of the image and using the corresponding density feature as supplementary information, the different regions with different densities in the image are assigned with different weights. So that the density feature of each region will change with the crowd density. With density feature, the crowd distribution can be described well and the crowd in different regions can be counted accurately.

The proposed network is similar to the CMTL. But in CMTL, it is only one-column CNN used for crowd counting feature extraction. The change in the scale of the crowd will decrease the counting performance of CMTL. So in the proposed network, the multi-column structure of MCNN is used to extract crowd counting feature. In order to get a more accurate density map to improve the accuracy of the counting results, the proposed method also improves the network structure. This part will be introduced in Section B.

Based on the above discussions, the details of the proposed architecture are presented in the following paragraphs. In the proposed network, two convolutional layers are used as the initial layers. The first convolutional layer has 16 convolutional kernels with a filter size of 9 × 9, the second convolutional layer has 32 convolutional kernels with a filter size of 7 × 7. The output features are processed with two sub-tasks.

The density classification sub-task is constructed by four convolutional layers, the number and size of convolutional kernels are represented by $16 \times 9 \times 9, 32 \times 7 \times 7, 16 \times 7 \times 7, 8 \times 7 \times 7$. These four convolutional layers are used for global features extraction, and the extracted global features are used for both density classification and crowd counting. After that, the extracted global features are fed back to adaptive max pooling layer and fully connected layer. The adaptive max pooling layer is used to get fixed size features. The last fully connected layer has five neurons for estimating the density level. Finally, the softmax classifier [31] is used for density classification.

In crowd counting sub-task, two-column CNN is used to process the output features from the initial convolutional layer. Then, the features extracted by crowd counting sub-task are concatenated with the global density features extracted by density classification sub-task. This structure is designed for obtaining final feature map. Then two convolutional layers are added to accomplish the final feature map. Two deconvolutional layers are used to restore the lost details in down-sampling process. Finally, a kind of 1 × 1 convolutional layer is used to process the output feature map in deconvolutional layers. The generated density map has the same size as the input image. It can provide more detailed feature information during the process of training. The parametric Rectified Linear Unit (PReLU) [32] is applied as activation function. Compared to ReLU in MCNN, PReLU can avoid the problems that functions will not be activated in the negative region. Its formula is expressed as follows:

$$f(x_i) = \begin{cases} x_i, & x_i > 0 \\ a_i x_i, & x_i \leq 0 \end{cases} \tag{1}$$

where, x_i is the input of nonlinear activation f on the i^{th} channel, and a_i is a coefficient which controls the slope of the negative part.

B. GENERATE A MORE COMPREHENSIVE DENSITY MAP

The accuracy of density maps is important for estimating the number of people. The comprehensive information contained in the density maps will affect the accuracy of the final estimation. In order to generate a more comprehensive density map, two aspects below are improved.

1) MAX-AVE POOLING

The features in the convolutional layer contain information related to spatial features, such as location and relative location. When the crowd distribution is even in the image, the max pooling will lose the relevant local spatial information. It will greatly influence the feature extraction and representation. In order to keep more local relevant information of the images, the proposed network uses the max-ave pooling to replace the max pooling. The max-ave pooling has been proven to the effective in image retrieval and image classification [33], [34].

The average pooling will keep the extracted features robust against small deformations and the max pooling can keep the extracted features unchanged [33]. The average pooling can be defined as (2) and the max pooling can be defined as (3):

$$F(v) = \frac{1}{T} \sum_{m=1}^{N} v_m \tag{2}$$

$$F(v) = \max_{1 \leq m \leq T} v_m \tag{3}$$

where, v_m represents the m^{th} pixel point extracted from T pixels in the sliding window of the image, m represents the spatial orientation of the element in the sliding window. The space pooling operator F defined above is used to map v_m to the corresponding statistical value.

The max-ave pooling combines the advantages of the above two pooling methods. The max-ave pooling is obtained by superimposing the average pooling and the max pooling with the same weight 1. It not only expands local receptive field of the image, but also keeps more accurate image feature information [33]. Its formula is defined as (4):

$$F(v) = \left(\frac{1}{T} \sum_{m=1}^{N} v_m \right) + \max_{1 \leq m \leq T} v_m \tag{4}$$

where, $F(v)$ represents the max-ave pooling obtained by superimposing the average pooling and the max pooling.

2) DECONVOLUTIONAL LAYER

During the calculation process in the pooling layers, the resolution of the feature map will be reduced. It will lead to loss of detailed feature information. In order to address this issue, the proposed network adds two deconvolutional layers [35].

The calculation methods of deconvolution and convolution are different. The deconvolution can be regarded as the reverse calculation of convolution. It is essentially an up-sampling process. During the calculation process in the pooling layers, the high-dimensional features will be down sampled to the low-dimensional features. It will reduce the resolution of the feature map. The position of the maximum in the low-dimensional feature matrix within each pooling region will be recorded in a position set. During the up-sampling process in the deconvolutional layers, the resolution of the feature map will be recovered. The high-dimensional feature matrix will be restructured and the maximum within each pooling region will be restored approximatively in the high-dimensional feature matrix according to the position set. The rest of the high-dimensional feature matrix will be set to zero [35].

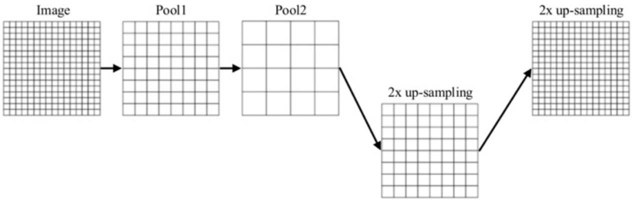


FIGURE 3. The down-sampling process in the convolution and the up-sampling process in the deconvolution.

The process of convolution down-sampling and deconvolution up-sampling is shown in Fig.3. In Fig.3, when the down-sampling operation in the first convolutional layer is finished, the resolution of output feature map in the pool1 becomes half of that in the input image. When the down-sampling operation of the second convolutional layer is finished, the resolution of output feature map in the pool2 becomes half of that in the pool1. That means it only keeps 1/4 resolution of the input image. Through adding two deconvolutional layers and four times up-sampling, the resolution of density map and lost details could be recovered approximatively. In Fig.3, the two up-samplings are performed by one deconvolutional layer. Another two up-samplings are performed by another deconvolutional layer. After two deconvolutional layers have been performed, the resolution and details of final feature map could be restored to the input image approximatively.

C. TRAINING LABLES

1) DENSITY MAP LABELS

Since the crowd density map can reflect the location information of the crowd, it is used as a training label for the crowd counting sub-task in this paper. The density map [37] of an image with N heads can be represented by the following formula (5):

$$F(X) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x) \tag{5}$$

where, $G_\sigma(x)$ is a two-dimensional Gaussian convolutional kernel. σ is width parameter. $\delta(x - x_i)$ represents the delta function, and x_i represents the position of head label.

The original image and the density map generated by formula (5) are shown in Fig.4. It can be observed that the crowd distribution in the ground truth density map is the same as that in the original image. Here, the number of people in the ground truth density map is 23 and it is exactly same with the original image.

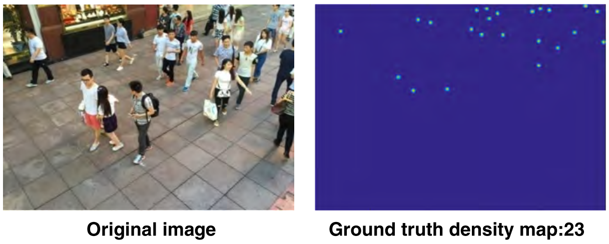


FIGURE 4. The first one is the original image. The second one is the ground truth density map.

2) DENSITY LEVEL CLASSIFICATION LABLES

The maximum number of people, $N_{\max}$, and the minimum number of people, $N_{\min}$, in the dataset are used to determine the range of the crowd number [18]. The number of density levels M is specified according to the scale of dataset. If the ground truth number of input image is GT, the density level can be determined by the following formula (6):

$$\text{Level} = \text{round}\left(\frac{GT}{\frac{N_{\max} - N_{\min}}{M}}\right) \tag{6}$$

where the denominator indicates the range of the crowd number in each density level, and round ()is used to round off the value. The density level number M in this experiment is set to 5.Then, the density level of each training sample is converted into a vector form.

D. NETWORK TRAINING

The proposed network is trained by cascading multi-task learning, and two sub-tasks need to be trained simultaneously.

For the density classification sub-task, the density level of each training sample is converted to the vector by the method proposed in Section C, and these vectors can be used to form the training set. The cross-entropy loss function [18], [23] is used for this sub-task. The formula is defined as (7):

$$L_1 = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[\left(p^i = j\right) F_1(X_i; \theta) \right] \tag{7}$$

where, N is the number of training images. M is the total density level classification number (M is signed a value of 5 in this paper). p^i is the ground truth classification level and X_i is the i^{th} training image. θ is the parameter learned from the convolutional network. Here, $F_1(X_i; \theta)$ is equal to $\log p_{i,j}$. It represents the probability that the i^{th} training image is estimated to be the j^{th} classification level. Since the loss function uses the error between the estimated density level and the ground truth density level for back propagation, the trained model can extract global density feature through the first four convolutional layers in the density classification sub-task.

For the crowd counting sub-task, the density map labels corresponding to each training sample are generated by the method proposed in Section C. The Euclidean distance is chosen as the loss function for this sub-task. It is defined as formula (8):

$$L_2 = \frac{1}{2N} \sum_{i=1}^{N} \left\| F_2(X_i; D_i; \theta) - G_i \right\|_2^2 \tag{8}$$

where, N represents the number of training samples, X_i represents the i^{th} training image, D_i represents the feature map obtained in the density classification sub-task, θ is the parameter learned from the convolutional network. $F_2(X_i; D_i; \theta)$ is the density map estimated by the crowd counting sub-task. G_i is the ground truth density map label and L_2 is the loss value between the estimated density map and the ground truth density map label. This loss function uses the error between the estimated density map and the ground truth density map for back propagation, so the trained model can extract the crowd features through crowd counting sub-task.

The whole network is jointly trained with combined loss function. The combined loss function L is defined as follows:

$$L = L_1 \sigma + L_2 \tag{9}$$

where, L_1 and L_2 are defined as above. σ is the weight (σ is signed a value of 0.001 in this paper).

IV. EXPERIMENT

This paper evaluated the performance of the proposed method on two different crowd counting datasets: the UCF_CC_50 and the ShanghaiTech. The whole network is trained and tested based on Pytorch framework. Hardware configuration is the GeForce GTX 1080Ti GPU with 11GB of memory and the CPU E5-2630.The operating environment is Ubuntu 16.04.3 with 32G RAM.

A. EVALUATION METRIC

The mean absolute error (MAE) and the mean squared error (MSE) are used to evaluate the performance on the test datasets. MAE represents the accuracy of the estimation and MSE indicates the robustness of the estimation. MAE and MSE are defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| z_i - z_i^{'} \right| \tag{10}$$

$$\text{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(z_i - z_i^{'}\right)^2} \tag{11}$$

where, N represents the total number of images in the test datasets, z_i and $z_i^{'}$ represent the ground truth value and the estimated value of the i^{th} image respectively.

The accuracy of obtained feature map determines the accuracy of estimated density map. The closer the estimated density map to the ground truth density map is, the more accurate of the obtained feature map is. So, this paper compares the estimated density map with the ground truth density map to test the accuracy of the obtained feature map. Two density maps are compared by the Peak Signal to Noise Ratio (PSNR) and the Structure Similarity Index Method [36] (SSIM). PSNR represents the distortion between estimated density map and ground truth density map. From three aspects: brightness, contrast, and structure, SSIM is used to measure the similarity between estimated density map and ground truth density map.

B. ABLATION EXPERIMENT

An ablation experiment is used to illustrate the effectiveness of the above method. The ShanghaiTech Part_A dataset is taken as the experimental testing dataset. During the experiments below, the performance of the crowd counting will be tested and analyzed by a step-by-step improvement method. The results are shown in Table 1 and Table 2.

TABLE 1. Comparison of estimation error under several different network configurations on Shanghaitech Part_A.

Method	MAE	MSE
MCNN [16]	110.2	173.2
MCNN + Global density feature	95.4	138.4
MCNN + Global density feature + Max-Ave Pooling	91.9	131.6
MCNN + Global density feature + Deconvolution	87.7	128.2
Ours	86.6	129.7

TABLE 2. Comparison of the accuracy of density maps under several different network configurations on Shanghaitech Part_A.

Method	PSNR	SSIM
MCNN [16]	20.38	0.549
MCNN + Global density feature	20.85	0.574
MCNN + Global density feature + Max-Ave Pooling	20.92	0.592
MCNN + Global density feature + Deconvolution	21.22	0.614
Ours	21.49	0.627

Table 1 shows that compared with MCNN, in the method with global density feature, the index MAE can be reduced 14.8 points and the index MSE can be reduced 34.8 points. In the method of using max-ave pooling with global density feature, the index MAE can be reduced 18.3 points and the index MSE can be reduced 41.6 points. In the method of adding deconvolutional layers with global density feature, the index MAE can be reduced 22.5 points and the index MSE can be reduced 45.0 points. In ours, that means using max-ave pooling and deconvolutional layers with global density feature, the index MAE can be reduced 23.6 points and the index MSE can be reduced 43.5 points. The result demonstrates that the method proposed effectively improves accuracy of crowd counting.

Table 2 shows that compared with MCNN, in the method with global density feature, the index PSNR can be increased 0.47 dB and the index SSIM can be increased 0.025 points. In the method of using max-ave pooling with global density feature, the index PSNR can be increased 0.54dB and the index SSIM can be increased 0.043 points. In the method of adding deconvolutional layers with global density feature, the index PSNR can be increased 0.84dB and the SSIM can be increased 0.065 points. In ours, that means using max-ave pooling and deconvolutional layers with global density feature, the index PSNR can be increased 1.11dB and the index SSIM can be increased 0.078 points. The result demonstrates that the proposed method can improve the accuracy of the estimated density map.

C. THE UCF_CC_50 DATASET

The UCF_CC_50 dataset is a high-density crowd dataset. This dataset contains 50 gray scale images of different scenes. The number of people per image ranges from 94 to 4543, and the average number of people per image is 1279.5. The five-fold cross validation can be employed to evaluate the results of the proposed method.

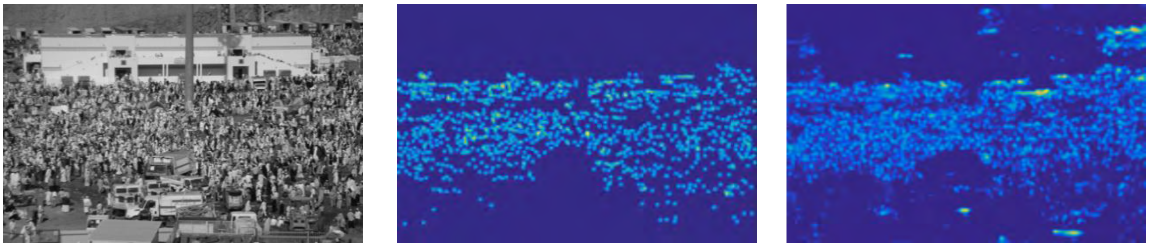
1) EXPERIMENTAL SETUP

In UCF_CC_50 dataset, the number of images available for training is limited and the number of people in per image is large. To ensure enough training images, data augmentation strategy is performed by randomly cropping 72 patches with size 225 × 225 from each image and flipping them. At the same time, noise points are randomly added to all patches. So the training dataset for verification will have 5760 images. The Adam optimization algorithm is used to minimize the cumulative error.

2) RESULTS

In this paper, the proposed method is compared with several recent methods on the UCF_CC_50 dataset. The results are shown in Table 3.

Table 3 shows that compared with the experimental results of MCNN, in ours, the index MAE is 70.9 points lower and the index MSE is 112.8 points lower. Compared with the experimental results of CMTL, in ours, the index MAE is 16.1 points lower and the MSE is 1.6 points lower. Compared with the experimental results of NetVLAD, in ours, the index MAE is 4.6 points lower and the MSE is 5.5 points lower. Compared with the experimental results of other methods in Table 3, the proposed method also has a significant



Original image

Ground truth density map:1050

Estimated density map:1088

FIGURE 5. The ground truth and estimated density map of testing images in UCF_CC_50 dataset.

TABLE 3. Comparison of MAE and MSE under several different methods on the UCF_CC_50 dataset.

Method	MAE	MSE
Lempitsky et al. [37]	493.4	487.1
Idrees et al. [8]	419.5	541.6
Zhang et al. [15]	467.0	498.5
CrowdNet [17]	452.5	-
MCNN [16]	377.6	509.1
Hydra2s [38]	333.7	425.3
Zeng et al. [39]	363.7	468.4
CMTL [18]	322.8	397.9
Switch-CNN [19]	318.1	439.2
TDF-CNN [40]	354.7	491.4
NetVLAD [20]	311.3	401.8
ComplexNet [41]	448.5	703.9
SaCNN [42]	314.9	439.2
MMCNN [23]	320.6	323.8
Ours	306.7	396.3

reduction in the index MAE and the index MSE. The result shows that the proposed method has higher accuracy and stability. Because of considering the different density features and changes in the scale, the performance of the proposed method has obvious improvement.

An experimental result is shown in Fig.5. The Original image is the original testing image with 1050 people. The Ground truth density map represents ground truth density map with 1050 ground truth people. The Estimated density map represents the estimated density map with 1088 estimated people. It can be seen that in the case of high density, the estimated crowd density map is similar to the ground truth density map, and the estimated number is close to ground truth number.

D. THE SHANGHAITECH DATASET

The ShanghaiTech dataset is divided to two parts: Part_A has 482 crowd images obtained from the Internet, the number of people in each image is between 33 and 3139, the average number of people per image is 501.4. Part_B contains 716 images captured from streets views, the number of each image is between 9 and 578, the average number of people per image is 123.6. The ShanghaiTech dataset has huge fluctuations on crowd distribution.

1) EXPERIMENTAL SETUP

In order to ensure enough training images, this paper randomly crops out 9 patches from each image and flips them, and each patch is 1/4 size of the original image. At the same time, noise points are randomly added to all patches. There are 482 images in the Part_A dataset, it contains 300 training images and 182 testing images. There are 716 images in the Part_B dataset, it contains 400 training images and 316 testing images. The training images of Part_A will increase to 5400 and the training images of part_B increase to 7200. The Adam optimization algorithm is used to minimize the cumulative error.

2) RESULTS

In this paper, the proposed method is compared with several recent methods on the ShanghaiTech dataset. The testing results are shown in Table 4 and Table 5.

TABLE 4. Comparison of MAE and MSE under several different methods on the Shanghaitech Part_A dataset.

Method	MAE	MSE
Zhang et al. [15]	181.8	277.7
MCNN [16]	110.2	173.2
CMTL [18]	101.3	152.4
Switch-CNN [19]	90.4	135.0
TDF-CNN [40]	97.5	145.1
NetVLAD [20]	107.6	169.3
CRCCNN [21]	107.0	162.1
SaCNN [42]	86.8	139.2
MMCNN [23]	91.2	128.6
Ours	86.6	129.7

In Table 4, the results in Part_A dataset show that compared with the experimental results of MCNN, in ours, the index

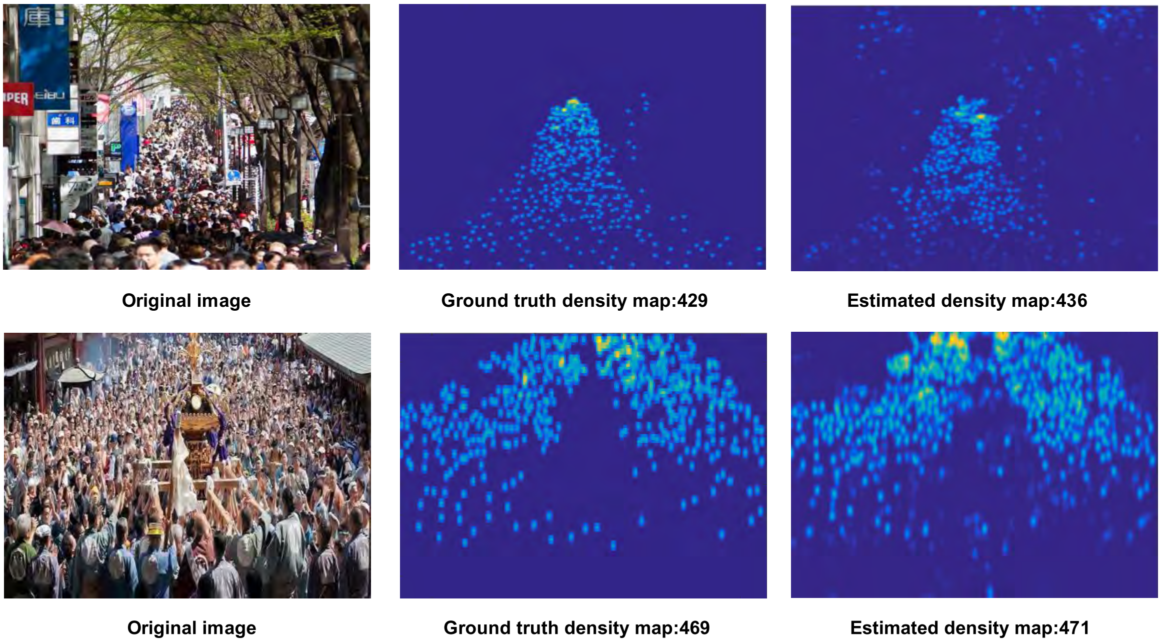

Original image
Ground truth density map:429
Estimated density map:436
Original image
Ground truth density map:469
Estimated density map:471

FIGURE 6. The ground truth and estimated density map of testing images in ShanghaiTech_PartA dataset.

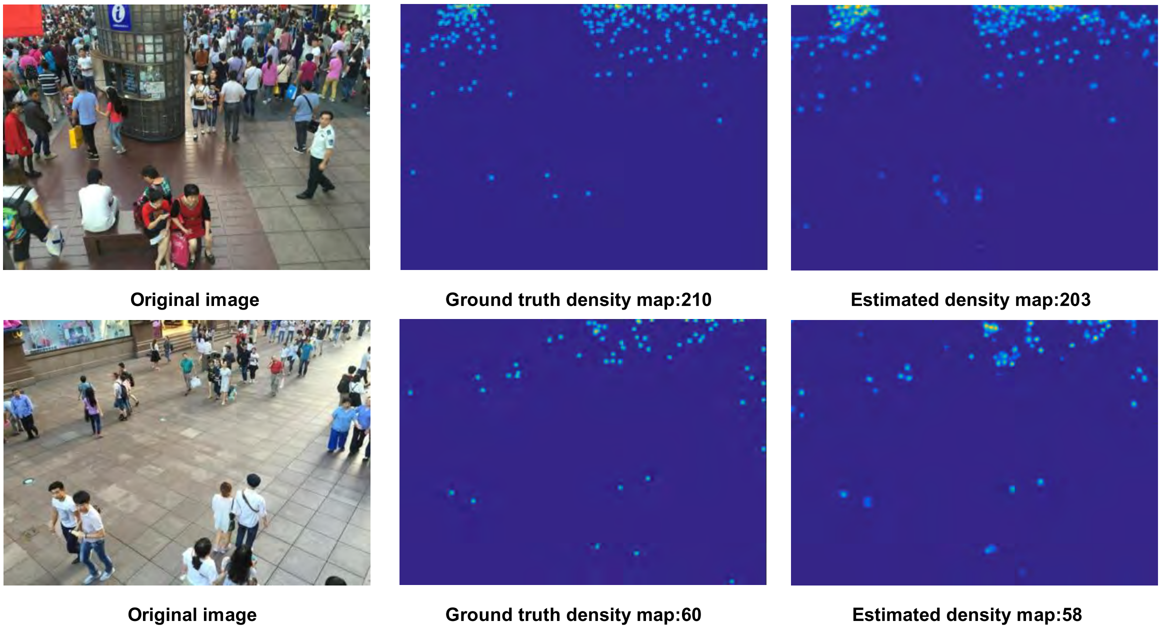

Original image
Ground truth density map:210
Estimated density map:203
Original image
Ground truth density map:60
Estimated density map:58

FIGURE 7. The ground truth and estimated density map of testing images in ShanghaiTech_PartB dataset.

MAE is 23.6 points lower and the index MSE is 43.5 points lower. Compared with the experimental results of CMTL, in ours, the index MAE is 14.7 points lower and the MSE is 22.7 points lower. Compared with the experimental results of SaCNN, in ours, the index MAE is 0.2 points lower and the index MSE is 9.5 points lower. Similar to the result in Table 3, the proposed method is superior to the other methods in Part_A dataset. The UCF_CC_50 dataset and the Part_A dataset both have extremely dense crowds. It means the proposed method has good performance when the crowd is extremely dense.

In Table 5, compared with the experimental results of MCNN in Part_B dataset, in ours, the index MAE is 7.1 points lower and the index MSE is 6.0 points lower. Compared with the experimental results of CMTL, in ours, the index MAE is 0.7 points lower. Compared with the experimental results of TDF-CNN and DecideNet, in ours, the index MAE is 1.4 points lower. SaCNN has the best performance in Part_B

TABLE 5. Comparison of MAE and MSE under several different methods on the Shanghaitech Part_B dataset.

Method	MAE	MSE
Zhang et al. [15]	32.0	49.8
MCNN [16]	26.4	41.3
CMTL [18]	20.0	31.1
Switch-CNN [19]	21.6	33.4
TDF-CNN [40]	20.7	32.8
NetVLAD [20]	21.4	33.9
DecideNet [22]	20.7	29.4
CRCCNN [21]	24.3	37.8
SaCNN [42]	16.2	25.8
MMCNN [23]	18.5	29.3
Ours	19.3	35.3

dataset due to its scale-adaptive strategy. The scale-adaptive strategy has good performance in low density crowd, such as Part_B dataset.

Some experimental results of the proposed method are shown in Fig.6 and Fig.7. Fig.6 shows the testing results of the two images on ShanghaiTech_PartA dataset. Fig.7 shows the testing results of the two images on ShanghaiTech_PartB dataset. These four images represent uneven crowd distributions. The results show that the proposed method is more accurate and robust compared with other methods on ShanghaiTech dataset.

V. CONCLUSION

In this paper, an improved convolutional neural network combined with global density feature is proposed. It is different from existing crowd counting methods. The proposed method focuses on uneven crowd distribution. Moreover, the max-ave pooling and deconvolutional layers are used to generate a more comprehensive density map. The experimental results show that the proposed method achieves competitive performance on different crowd datasets. Due to the high density crowd, some backgrounds will be taken as people by mistakes. It will bring about noise in the estimated density map and influence the counting results. For the future work, we will focus on reducing the noise in the estimated density map and improving the accuracy of counting.

REFERENCES

[1] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, ‘‘Estimation of number of people in crowded scenes using perspective transformation,’’ *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 6, pp. 645–654, Nov. 2001.

[2] P. Gardzinski, K. Kowalak, L. Kaminski, and S. Mackowiak, ‘‘Crowd density estimation based on voxel model in multi-view surveillance systems,’’ in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Sep. 2015, pp. 216–219.

[3] M. Li, Z. Zhang, K. Huang, and T. Tan, ‘‘Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection,’’ in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[4] T. Zhao, R. Nevatia, and B. Wu, ‘‘Segmentation and tracking of multiple humans in crowded environments,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.

[5] X. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, ‘‘LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data,’’ *World Wide Web*, vol. 21, no. 3, pp. 825–847, May 2018.

[6] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, ‘‘Privacy preserving crowd monitoring: Counting people without people models or tracking,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[7] A. B. Chan and N. Vasconcelos, ‘‘Counting people with low-level features and Bayesian regression,’’ *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2160–2177, Apr. 2012.

[8] H. Idrees, I. Saleem, C. Seibert, and S. Mubarak, ‘‘Multi-source multi-scale counting in extremely dense crowd images,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[9] W. Ma, L. Huang, and C. Liu, ‘‘Crowd density analysis using co-occurrence texture features,’’ in *Proc. 5th Int. Conf. Comput. Sci. Converg. Inf. Technol.*, Nov./Dec. 2010, pp. 170–175.

[10] D. Kong, D. Gray, and H. Tao, ‘‘A viewpoint invariant approach for crowd counting,’’ in *Proc. 18th Int. Conf. Pattern Recognit.*, Aug. 2006, pp. 1187–1190.

[11] X. Kong, F. Xia, Z. Ning, A. Rahim, Y. Cai, Z. Gao, and J. Ma, ‘‘Mobility dataset generation for vehicular social networks based on floating car data,’’ *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3874–3886, May 2018.

[12] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, ‘‘Time-location-relationship combined service recommendation based on taxi trajectory data,’’ *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[13] J. Zhang, P. Liu, F. Zhang, and Q. Song, ‘‘CloudNet: Ground-based cloud classification with deep convolutional neural network,’’ *Geophys. Res. Lett.*, vol. 45, no. 16, pp. 8665–8672, Aug. 2018.

[14] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, ‘‘Salient object segmentation via effective integration of saliency and objectness,’’ *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, Aug. 2017.

[15] C. Zhang, H. Li, X. Wang, and X. Yang, ‘‘Cross-scene crowd counting via deep convolutional neural networks,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 833–841.

[16] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, ‘‘Single-image crowd counting via multi-column convolutional neural network,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 589–597.

[17] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, ‘‘CrowdNet: A deep convolutional network for dense crowd counting,’’ in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 640–644.

[18] V. A. Sindagi and V. M. Patel, ‘‘CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting,’’ in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug./Sep. 2017, pp. 1–6.

[19] D. B. Sam, S. Surya, and R. V. Babu, ‘‘Switching convolutional neural network for crowd counting,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4031–4039.

[20] Z. Shi, L. Zhang, Y. Sun, and Y. Ye, ‘‘Multiscale multitask deep NetVLAD for crowd counting,’’ *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4953–4962, Nov. 2018.

[21] J. Fu, H. Yang, P. Liu, and Y. Hu, ‘‘A CNN-RNN neural network join long short-term memory for crowd counting and density estimation,’’ in *Proc. IEEE Int. Conf. Adv. Manuf.*, Nov. 2018, pp. 471–474.

[22] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, ‘‘DecideNet: Counting varying density crowds through attention guided detection and density estimation,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5197–5206.

[23] B. Yang, J. Cao, N. Wang, Y. Zhang, and L. Zou, ‘‘Counting challenging crowds robustly using a multi-column multi-task convolutional neural network,’’ *Signal Process., Image Commun.*, vol. 64, pp. 118–129, Mar. 2018.

[24] V. A. Sindagi and V. M. Patel, ‘‘Generating high-quality crowd density maps using contextual pyramid CNNs,’’ in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1861–1870.

[25] P. Chen, Y. Tang, L. Wang, and X. He, ‘‘Crowd density estimation based on multi-level feature fusion,’’ *J. Image, Graph.*, vol. 23, no. 8, pp. 1181–1192, Aug. 2018.

[26] J. Dai, K. He, and J. Sun, ‘‘Instance-aware semantic segmentation via multi-task network cascades,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3150–3158.

[27] P. Wang, G. Tian, and H. Chen, "A saliency detection model combined local and global features," in *Proc. Chin. Automat. Congr. (CAC)*, Oct. 2017, pp. 2863–2870.

[28] X. Tan, H. Zhu, Z. Shao, X. Hou, Y. Hao, and L. Ma, "Saliency detection by deep network with boundary refinement and global context," in *Proc. IEEE Int. Conf. Multimedia, Expo*, Jul. 2018, pp. 1–6.

[29] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Scene parsing with global context embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2650–2658.

[30] H. Fradi, X. Zhao, and J.-L. Dugelay, "Crowd density analysis using subspace learning on local binary pattern," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2013, pp. 1–6.

[31] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, Architectures and Applications* (NATO ASI Series), vol. 68, F. F. Soulié and J. Hérault, Eds. 1990, pp. 227–236.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[33] C. Bai, L. Huang, J.-N. Chen, X. Pan, and S.-Y. Chen, "Optimization of deep convolutional neural network for large scale image classification," *J. Softw.*, vol. 29, no. 4, pp. 1029–1038, Jan. 2018.

[34] C. Bai, L. Huang, X. Pan, J. Zheng, and S. Chen, "Optimization of deep convolutional neural network for large scale image retrieval," *Neurocomputing*, vol. 303, pp. 60–67, Aug. 2018.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 818–833.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[37] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2010, pp. 1324–1332.

[38] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 615–629.

[39] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 465–469.

[40] D. B. Sam and R. V. Babu, "Top-down feedback for crowd counting convolutional neural network," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 7323–7330.

[41] M. Matlacz and G. Sarwas, "Crowd counting using complex convolutional neural network," in *Proc. Signal Process., Algorithms, Archit., Arrangements, Appl. (SPA)*, Sep. 2018, pp. 88–92.

[42] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1113–1121.

ZHI LIU received the B.S. degree in automatic control and the M.S. degree in system engineering from Xi'an Jiaotong University, Xi'an, China, in 1991 and 1994, respectively, and the Ph.D. degree in computer application from Zhejiang University, Hangzhou, China, in 2001. She is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou. She is a member of the China Computer Federation. Her research interests mainly include 3D model retrieval, image processing, and intelligent transportation systems.

YUE CHEN received the B.S. degree in computer science and technology from Zhejiang Sci-Tech University, Hangzhou, China, in 2016. He is currently pursuing the M.S. degree in computer technology with the Zhejiang University of Technology, Hangzhou. His research interests include image processing and artificial intelligence.

BO CHEN received the B.S. degree in computer science and the M.S. degree in computer science and technology from Hangzhou University, Hangzhou, China, in 1992 and 1995, respectively. He is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou. His research interests include artificial intelligence and image processing.

LINAN ZHU received the Ph.D. degree in mechanical manufacturing and automation from the Zhejiang University of Technology, in 2014, where he is currently an Associate Professor with the College of Computer Science and Technology. His research interests include cloud manufacturing, resource scheduling, and artificial intelligence.

DU WU is currently pursuing the B.S. degree in computer science and technology with Jilin University, Changchun, China. His research interests include image processing and artificial intelligence.

GUOJIANG SHEN received the B.Sc. degree in control theory and control engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1999 and 2004, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou. His current research interests include artificial intelligence, big data analytics, and intelligent transportation systems.