



# Crowd density estimation based on multi scale features fusion network with reverse attention mechanism

Yong-Chao Li<sup>1</sup> · Rui-Sheng Jia<sup>1,2</sup> · Ying-Xiang Hu<sup>1</sup> · Dong-Nuo Han<sup>1</sup> · Hong-Mei Sun<sup>1,2</sup>

Accepted: 4 January 2022 / Published online: 22 February 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

Deep learning has made substantial progress in crowd counting, but in practical applications, due to interference factors such as perspective distortion and complex background, the existing methods still have large errors in counting. In response to the above problems, this paper designs a multi-scale feature fusion network (IA-MFFCN) based on the reverse attention mechanism, which maps the image to the crowd density map for counting. The network consists of three parts: feature extraction module, inverse attention module, and back-end module. First, to overcome the problem of perspective distortion, deeper single-column CNNs was designed as a feature extraction module to extract multi-scale feature information and merge them; second, to avoid interference of complex backgrounds, the inverse attention module was designed, through the multi-scale inverse attention mechanism, reducing the influence of noise on counting accuracy. Finally, to generate a high-quality crowd density map, dilation convolution was introduced. Simultaneously, to enhance the sensitivity of the network to crowd counting, a comprehensive loss function based on Euclidean loss and predicted population loss is designed to improve training accuracy, to produce a more accurate density value. Experiments show that compared with the comparison algorithm, the algorithm in this paper has a significant reduction in the mean absolute error (MAE) and mean square error (MSE) on the ShanghaiTech dataset, UCF\_CC\_50 dataset and WorldExpo`10 dataset.

**Keywords** Crowd density estimation · Multi-scale feature fusion · Inverse attention mechanism · Dilated convolution · High-quality density map

## 1 Introduction

At present, with the development of social economy and the rapid growth of urban population(as shown in Fig. 1(a)), the problem of urban public safety becomes increasingly important. Due to the excessive crowd density caused by aggregation behavior, serious stampedes occurred on the Bund and Mecca pilgrimage in Shanghai, caused serious group safety accident [1], Therefore, it is of great significance for social

public safety to estimate crowd density through surveillance video images and prevent the occurrence of overcrowding and abnormal situations.

Early crowd density estimation methods can be roughly divided into detection-based and regression-based. These detection-based methods use the target detector [2] to detect every pedestrian in the crowd and count the detection results to obtain the final number of people(as shown in Fig. 1(b)). This kind of method is suitable for sparse crowd scenes, but it is difficult to play a role in crowded scenes with occlusion phenomena. These based-regression methods avoid counting directly through target detection, and realize counting by learning the mapping relationship between image features and corresponding number of people [3] (as shown in Fig. 1(c)). These based-regression methods alleviate the effect of occlusion phenomenon. However, the effect of this kind of method on feature extraction limits its performance to a great extent, and it is difficult to play a good role in the highly crowded crowd scene. In recent years, with the development of deep learning[4, 5], the method of convolutional

---

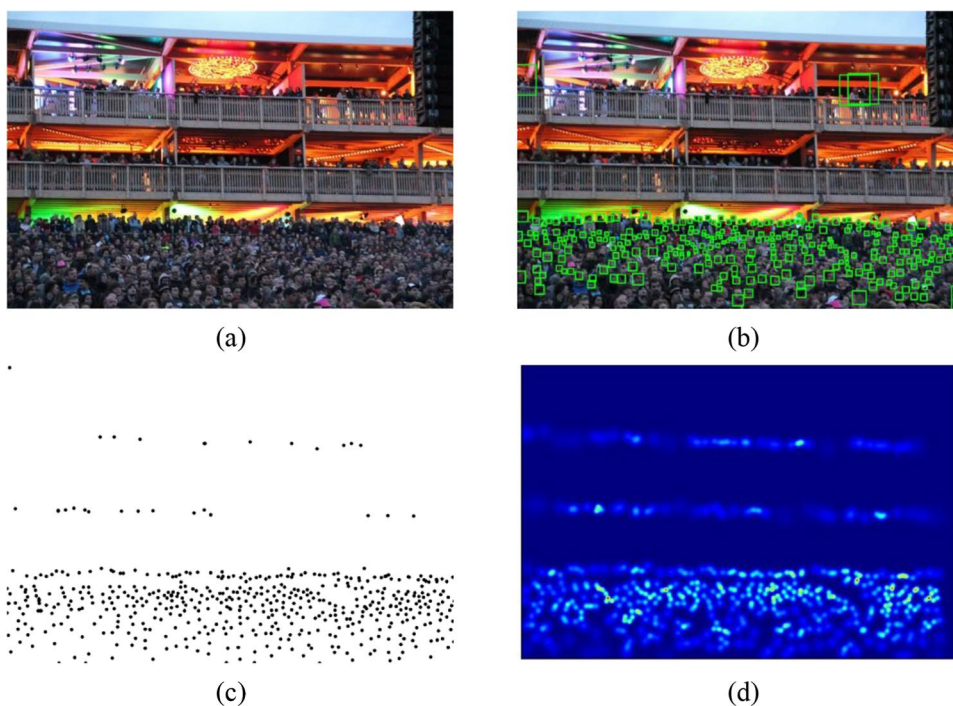
✉ Rui-Sheng Jia  
jrs716@163.com

✉ Hong-Mei Sun  
shm0221@163.com

<sup>1</sup> College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

<sup>2</sup> Shandong Province Key Laboratory of Wisdom Mine Information Technology, Shandong University of Science and Technology, Qingdao 266590, China

**Fig. 1** Comparison between traditional methods and deep learning methods. (a) Large-scale dense crowd images; (b) Results based on detection methods; (c) Results based on regression methods; (d) Results based on deep learning methods



neural networks has been widely used in the research of crowd density estimation [6–12]. This kind of method is not to calculate the number of people directly, but to predict a crowd density map by convolution neural network (as shown in Fig. 1 (d)), the pixel value of the image reflects the crowd density of this position in the original image. The number of people corresponding to the original image can be obtained using pixel accumulation of the density map, which is suitable for the crowd in complex scenes and improves the accuracy of crowd density estimation to a certain extent. At present, the main challenges of crowd density estimation are as follows:

- (1) Due to the perspective distortion problem caused by different shooting distance and angle of the monitoring camera, there may be people and heads with different scales in the same picture. Although multi-column convolution network has made great progress in solving the perspective distortion problem [13], it also has some obvious defects such as extracting redundant features [14], moreover, when the network becomes complex, it will increase the network computing resources.
- (2) Due to background interference, there may be a large deviation between the crowd density map and the Ground truth, which leads to a large deviation between the final estimated number and the actual number.

In summary, although the existing methods can adapt to problems such as occlusion and uneven distribution to a certain extent through autonomous learning. However, the problems caused by the interference of the complex

background in the crowd and perspective distortion are still not well solved, and almost all methods are to add an additional information or task to enhance the single crowd-counting task. Therefore, in response to the above problems, this paper proposes a multi-scale feature fusion network based on inverse attention mechanism (IA-MFFCN). This article contributes as follows:

- (1) In order to overcome multiple columns of CNNs extracting redundant features and wasting training resources, this paper designs a deeper single column CNNs as a feature extraction module, which can extract low-level and high-level features, and merge them to achieve different scale features extract;
- (2) To avoid affecting the accuracy of crowd counting due to background interference, a reverse attention module [15–19] is used to process low-level features containing background noise, separate the crowd and background information, and reduce the interference of background on crowd counting.
- (3) To improve the quality of the generated crowd density map, a set of dilated convolutions is introduced to expand the receptive field, so that the output contains more detailed spatial and global information.
- (4) To enhance the network's sensitivity to crowd counting in high-density scenarios, a new comprehensive loss function is designed. This loss function adds a loss based on the number of regression people to the traditional Euclidean loss function to better measure the training performance. The gap between the estimated value and the true value produces a more accurate density value and improves training accuracy.

In the open datasets (ShanghaiTech dataset, UCF\_CC\_50 dataset and WorldExpo`10 dataset), the experimental results show that the performance of the method is better than other comparison methods.

## 2 Related work

A single image crowd density estimation has been developed for many years, because it plays an important role in the field of public safety, so it has been a long-term concern of researchers in the field of computer vision. However, due to occlusion, perspective distortion, background interference, uneven distribution and so on, the improvement of counting accuracy is still facing considerable challenges. The existing methods can be divided into three categories: detection-based, regression-based and CNN-based. Detection-based and regression-based methods belong to traditional counting methods. In recent years, CNN-based methods have shown better performance than traditional methods in crowd-counting task.

**Detection-based methods** Most of the early studies are focused on the detection-based methods, which use a detector similar to moving window to detect the people in the image, and then calculate the number of people using statistics (as shown in Fig. 1 (b)). This kind of method should train a detector to extract the features of human head or body from the image, the cost of training this detector is colossal, and it needs to scan the image globally, in the past studies, many kinds of detectors have been used to detect individuals in the image, but the effect is not superb. In 2001, Lin et al. [19] used Haar wavelet classifier to detect the head characteristics of the population. In 2005, Navneet Dalal et al. [20] replaced Haar wavelet classifier with HOG (histogram oriented gradient) classifier to detect head features of people. In 2008, Li et al. [21] proposed a head shoulder detector with foreground segmentation. In 2009, Pedro F Felzenszwalb et al. [22] tried to detect some typical features of the body, rather than the whole body, because in crowd scenes, the human body is always covered. In 2011, Dollar et al. [23] used a detector similar to a moving window to detect the human body and calculate the number of people in the image. However, detection-based counting is limited by the occlusion between people in the crowded scene. As the crowd becomes dense, the detection performance decreases rapidly.

**Regression-Based methods** Because the detection-based methods can't adapt to the scene of high crowd density estimation, the regression-based methods are proposed. The regression-based methods directly map between the image features and the crowd density estimation, which avoid the complex

process of training the detector. These methods mainly have two steps: First, extract the feature information of the crowd image, including head for body features, foreground features, texture features and so on, these features are used to generate low-level information; Second, use regression function to calculate the number of people (as shown in Fig. 1 (c)), the common regression methods are ridge regression [24], Bayesian Poisson regression [25], random forest [26], etc. In 2009, Chan et al. [25] used foreground and texture features to generate low-level information, and calculated the number of people after learning the relationship between the crowd and the extracted features. In 2010, Lempitsky et al. [27] proposed a density map estimation method, which uses linear mapping between the features in a local region and its density map to integrate saliency problems. In 2012, Chan et al. [28] proposed the Bayesian model of discrete regression for crowd density estimation, which connects the proposed approximate Bayesian Poisson regression with Gaussian kernel to realize the crowd density estimation from low-level features. In 2013, Idrees et al. [3] proposed a model, which introduced Fourier analysis and SIFT (Scale invariant feature transform) to extract foreground and texture features of interested points to complete the crowd density estimation. In 2015, Pham et al. [26], to solve the gain problem of ideal linear mapping, proposed using random forest regression to learn nonlinear mapping instead of linear mapping. Regression-based methods are effective in some specific scenarios, such as low-density crowd or fixed monitoring perspective. However, for the presence of perspective distortion and dense crowd, the counting ability is insufficient.

**CNN-Based methods** With the development of deep learning, CNN is widely used in the field of crowd density estimation, and has shown considerable advantages. In 2015, Zhang et al. [29] applied convolutional neural networks to crowd density estimation for the first time, they trained convolutional neural networks to regress the crowd density map (as shown in Fig. 1 (d)), used density and perspective information to get the image again, and then used the trained network to predict the density map. However, their method needs perspective view during the training process and the testing process, which is difficult to achieve in the practical application of crowd density estimation. In 2016, Zhang et al. [13] proposed an improved structure, using multi-scale convolutional neural networks (MCNN) architecture to deal with large-scale changes in crowd scenes, and using  $1 \times 1$  convolution kernel to fuse data from each feature map trained by the convolutional network of a specific size is returned to the density maps, however, the interference of complex background affects the accuracy of counting. In 2017, Sindagi et al. [30] proposed an end-to-end cascaded multi task learning (CMTL), which uses the cascaded neural network to jointly learn the crowd density estimation classification and density map estimation to solve the problem

of uneven proportion crowd. By learning, the crowd density estimation is divided into different groups, and a high-level prior is combined into the density map estimation, it solves the problem of uneven crowd distribution. However, in the process of training, this method relies too much on the classifier, and takes a long time to the pre-training the network before training the classifier. At the same time, it also ignores the background interference. In 2018, Li et al. [14] proposed a crowd density estimation structure in congestion scenarios (CSRNet), which effectively expanded the receptive field of the network and generated high-quality density map by using dilated convolution, this is a great improvement. However, this method cannot extract multi-scale information, resulting in the loss of some crowd features, and the complex background interferes with counting accuracy. In 2019, Liu et al. [31] proposed the context area network (CANet), which uses an end-to-end deep structure that can be trained. It combines the features obtained by using multiple receptive fields, and learns the importance of the features in each position of the image, to solve the problem of crowd perspective distortion. However, due to the large amount of background noise in the crowd image, it is difficult to separate the crowd features from the background features, which leads to complex background noise seriously affects the counting accuracy. In 2019, Wang et al. [32] constructed a large-scale and diversified synthetic crowd density estimation dataset to the pre-training the spatial full convolution network they designed, which solved the problems of large-scale crowd changes, and different degrees of over fitting due to missing data in the changeable environment, but the background noise interference still affected the accuracy of counting. Almost all methods are to add additional information or task to enhance the single population counting task. In 2020, Zhu et al. [33] proposed multi-scale, multi-level characteristics of polymer network (MFANet), solve the network cannot accept a large domain feature-extracting figure, through the accurate and efficient end-to-end training to crowd counting. However, the mutual occlusion between crowds and the interference of complex environment still limit the accuracy of the count. In 2020, Liu et al. [34] proposed a crowd counting method based on self-attention mechanism, some of the sampling

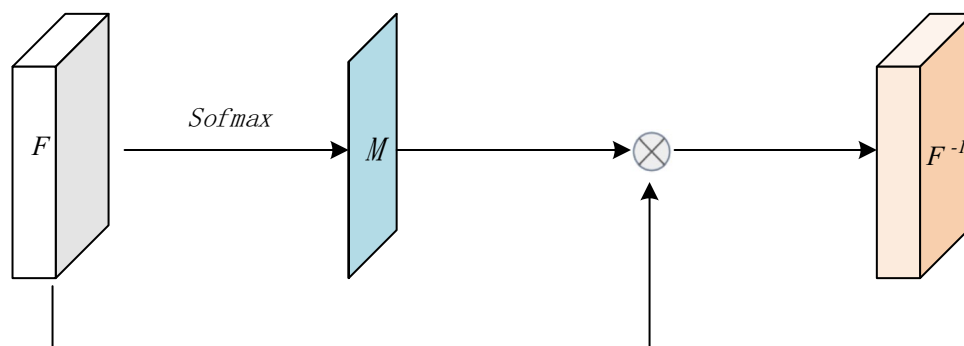
positions are shifted to the occluded crowd by shifting the sampling points, the problem of crowd occlusion is solved, but there is still a perspective distortion problem that interferes with the accuracy of this counting. In 2021, Gu et al. [35] present an innovative Context-Aware Pyramid Attention Network for crowd counting which is realized by extracting rich contextual features and dealing with dependencies on space and channels, but this work was not considered that the extracted features are too rich, and interference factors will be extracted, which will affect the accuracy of counting.

**Attention mechanism crowd counting** With the application of density map mapping in the field of crowd counting, this work has achieved counting while obtaining detailed information about the population distribution. Therefore, the researcher expects that the value of each head position area in the output density map adds to 1 and the value of the background part is 0, which has higher requirements for filtering of the background and the prominent position of the human head. Therefore, inspired by the visual attention mechanism of the human brain in the cognitive process that can focus on processing the information obtained from the visual cortex, in 2015, Richard et al. [36] proposed an attention mechanism expressed by a mathematical model to automatically generate image analysis sentences for images. After continuous development, the attention convolution module and its calculation process have evolved into the structure shown in Fig. 2.

As shown in Fig. 2,  $F$  is the combination of feature maps extracted by the convolutional neural network. After the processing of the Sigmoid function, the attention map  $M$  is obtained, extend  $M$  to the same number of channels as  $F$  and perform a channel-by-channel Hadamard product with it to obtain the attention feature  $F^{-1}$ , and then proceed to follow-up processing. The purpose of the attention module is to learn the different degrees of attention corresponding to different positions in the space, that is, the size of the weight. Taking the settings in Fig. 2 as an example, the attention map is generated by the following formula:

$$S = \varphi(W \odot F + b) \quad (1)$$

Fig. 2 Attention mechanism





$$M_P = \frac{e^{s_P}}{\sum_{P' \in P} e^{s_{P'}}} \quad (2)$$

Among them,  $\odot$  represents a convolution operation,  $P$  represents a different position in  $M$ , and  $M_P$  reflects the importance of the feature of the position  $P$  to the task, that is, the weight corresponding to the point  $P$ .

$$F^{-1} = F \otimes M \quad (3)$$

The obtained attention map  $M$  and feature map  $F$  are processed as shown in formula (3), and then the attention feature  $F_{\cdot j}$  is obtained.

### 3 Proposed solution

#### 3.1 Network architecture

To solve the problem of perspective distortion and complex background interference in large-scale dense crowd, a multi-scale feature fusion network based on inverse attention mechanism is designed. The network structure of this paper is shown in Fig. 3.

The network is composed of three modules. First, a feature extraction module is designed, which extracts features by constructing a deeper single-column CNNs, and uses low-level feature extraction networks and high-level feature extraction networks to extract feature information of different scales, and merge them, to solve the problem of perspective distortion in crowd counting; second, a inverse attention module is designed to reduce the noise of the feature information extracted by the low-level extraction network, separate crowd information and background information, and reduce the interference of background on counting to solve the interference problem of the complex background in crowd counting; finally, the dilated convolution is introduced as the back-end module to expand the receptive field of the network, and the output contains more detailed spatial and global information, which improves the quality of the generated crowd density map and improves the accuracy of the estimated number of people.

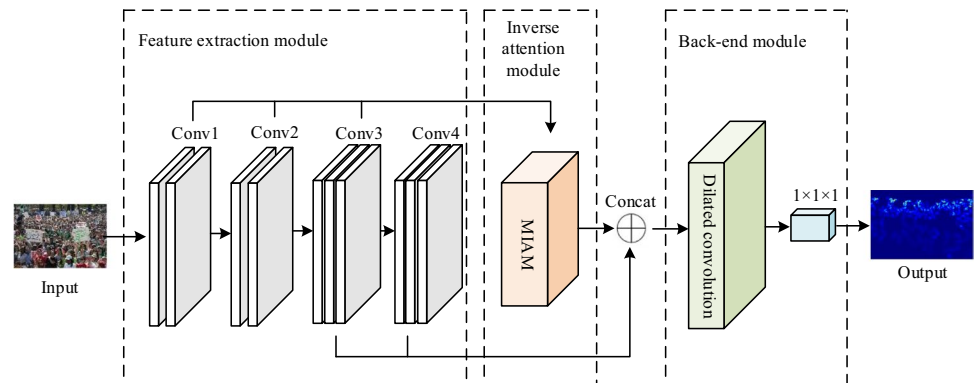
#### 3.2 Feature extraction module

In large-scale dense crowd, perspective distortion caused by different distances between people and cameras in the same scene will lead to the diversity of human head scale. To extract the crowd features of different scales and solve the perspective distortion problem of large-scale dense crowd, the feature extraction module uses multi-scale feature fusion network with single column convolution kernel. As shown in Fig. 3, the feature extraction module is composed of four convolution blocks: Conv1, Conv2, Conv3, Conv4. After each convolution block, the maximum -pooling layer with step size of 2 is used. Conv1 consists of 2 layers  $3 \times 3 \times 64$  convolution kernel, Conv2 consists of 2 layers  $3 \times 3 \times 128$  convolution kernel, Conv3 consists of 3 layers  $3 \times 3 \times 256$  convolution kernel, Conv4 consists of 3 layers  $3 \times 3 \times 512$  convolution kernel; Conv1 and Conv2 can extract the edge information and low pixel level head spots, Conv3 and Conv4 can extract the internal details of the crowd, face and body information. These different level feature maps have multi-scale features to avoid inaccurate counting due to incomplete feature information. Through the maximum-pooling layer, the corresponding sizes of these output feature images are 1/2, 1/4, 1/8 and 1/16 of the original input image respectively. Through up-sampling, the feature maps output by Conv2 and Conv3 are restored to 1/2 of the original image size, the output feature images of Conv1, Conv2 and Conv3 are superimposed to obtain the crowd margin information feature image  $F_{margin}$ ; Through up-sampling, the feature maps output by Conv4 are restored to 1/8 of the original image size, the output feature images of Conv3 and Conv4 are superimposed to obtain the crowd inside information feature image  $F_{inside}$ .

#### 3.3 Inverse attention module

When feature extraction module Conv1, Conv2, Conv3 convolution block extracts features, background interference information may be extracted as the crowd, which may lead to complex background interference, thus affecting the accuracy of counting. To solve the problem of complex background interference,

**Fig. 3** Network structure diagram of IA-MFFCN



this paper introduces the inverse attention module to distinguish the crowd from background noise. Inverse attention block (IAB) proposed by Sindagi and Patelp [37] solves the challenging problem of crowd density estimation in crowded scenes. It effectively injects segmentation information into counting network through inverse attention block, which leads to significant improvement, however, it uses a column of convolution layers to extract features, which has limited ability to extract features, here we design a module "multi-column inverse attention module" (MIAM) to improve the ability of feature information extracted by the module, and use it to solve the interference problem of complex background. MIAM is shown in Fig. 4.

The output feature maps of Conv1, Conv2, and Conv3 are superimposed to obtain the crowd edge information feature map  $F_{margin}$  as the input of the inverse attention module, and the  $F_{margin}$  is extracted through the multi-column convolution module (MCB), and the MCB uses two-column convolution to extract features, the network structure of the first column is: Conv<sub>448,128,1</sub>—ReLU—Conv<sub>128,64,3</sub>—ReLU—Max pooling—Conv<sub>64,32,1</sub>; the network structure of the second column is: Conv<sub>128,32,3</sub>-ReLU-Max pooling-Conv<sub>32,32,1</sub>, superimpose the feature maps of the output of the two columns of networks, and finally use a Conv<sub>64,32,1</sub>-Ave pooling convolution block to fuse the features extracted by the two columns of networks, where Conv<sub>I,O,K</sub>: I is the number of input feature map channels, O is the number of output feature map channels, K is the size of the convolution kernel, ReLU(Rectified Linear Unit) is the ReLU function, Max pooling is the maximum-pooling layer with a step size of 2, Ave pooling is the Average-pooling layer with a step size of 2. On the basis of IAB, the width of the network is expanded, a network branch is added, and the ability of extracting feature information is optimized; then obtain the background area ( $A^{-1}$ ) of the input feature image through the Sigmoid function,  $A^{-1}$  and  $F_{margin}$  will be weighted, that is  $F_{margin} \otimes A^{-1}$ , and then subtracted it from  $F_{margin}$  to reduce the influence of background on crowd density estimation. Its calculation formula is as follows:

$$F_{margin}^{-1} = F_{margin} - F_{margin} \otimes A^{-1} \tag{4}$$

The result in the formula is a feature map that suppresses background interference.

### 3.4 Back-end module

The feature extraction module Conv3 and Conv4 output feature maps are superimposed to obtain the inside feature map of the crowd  $F_{inside}$  is 1/8 of the original image size, and the  $F_{margin}^{-1}$  output by the inverse attention module is 1/8 of the original image size, so the two feature maps are superimposed to obtain  $F_{add}$  and sent to the back-end module. After the feature map passes through the pooling layer, although the over-fitting can be controlled and the invariance is maintained, the spatial resolution is reduced, part of the spatial information is lost, and the quality of the density map is not high enough. Therefore, we use a dilated convolution in the back-end module. As shown in Fig. 5, it consists of 6 layers of convolutional layers with the dilation rate of 2 and a layer of  $1 \times 1$  convolutional layers. The number of channels in each convolutional layer is different. The specific parameters are Conv<sub>3,512,2</sub>—Conv<sub>3,512,2</sub>—Conv<sub>3,256,2</sub>—Conv<sub>3,128,2</sub>—Conv<sub>3,64,2</sub>, where Conv K, N, R: K represents the size of the convolution kernel, N represents the number of channels, and R represents the dilation parameter;

The two-dimensional dilation convolution is defined as follows:

$$F_{end}(m, n) = \sum_{i=1}^M \sum_{j=1}^N F_{add}(m + r \times i, n + r \times j)w(i, j) \tag{5}$$

In the formula,  $F_{end}(m, n)$  is the output of dilated convolution from input  $F_{add}(m, n)$  and a filter  $w(i, j)$  with the length and the width of  $M$  and  $N$  respectively,  $r$  is the dilation parameter of dilated convolution. Send  $F_{end}$  to the  $1 \times 1 \times 1$  convolutional layer, and finally, the network output a high-quality crowd density map.

Fig. 4 Multi-column inverse attention module

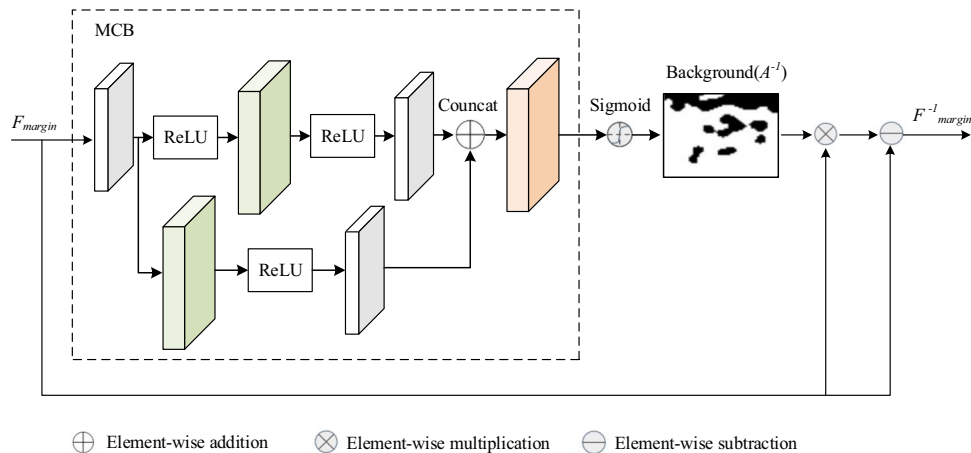
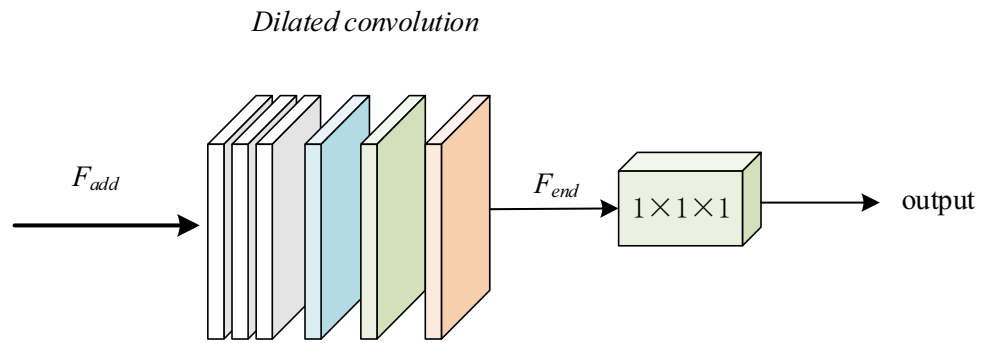


Fig. 5 Back-end module



### 3.5 Loss function

The existing mainstream algorithms all set the Euclidean loss on the pixel as the loss function in the training process, but the loss function does not take into account the global count error of the input image. Therefore, we have designed a comprehensive loss function, which includes two parts: the Euclidean loss function and the loss function based on the predicted number of crowd. The Euclidean loss function is as follows:

$$L_1 = \frac{1}{2N} \sum_{i=1}^N \|Z(X_i, \theta) - Z_i^{GT}\|_2^2 \tag{6}$$

In the formula,  $N$  is the number of pictures in the training set,  $X_i$  is the input picture,  $\theta$  is a set of parameters that can be learned,  $Z(X_i, q)$  represents the estimated density of the image  $i$ ,  $Z_i^{GT}$  represents the ground truth of the image  $i$ .

The loss function based on the predicted number of people is as follows:

$$L_2 = \|\hat{C} - C\|^2 \tag{7}$$

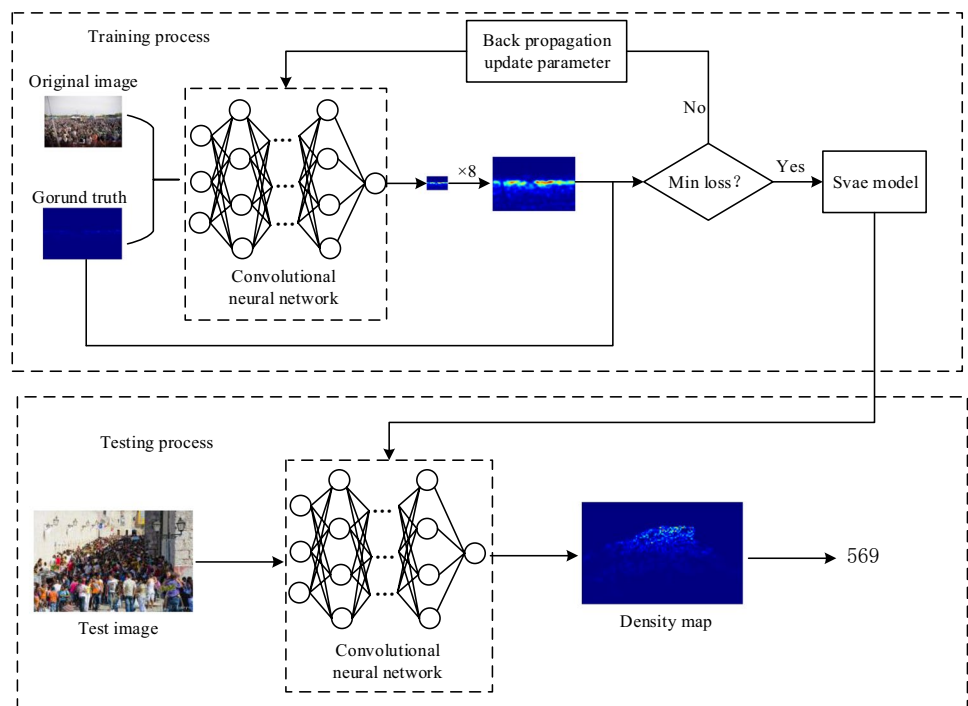
In the formula,  $\hat{C}$  represents the estimated number of crowd, and  $C$  represents the real number of crowd.

The comprehensive loss function is as follows:

$$L = \alpha L_1 + (1 - \alpha) L_2 \tag{8}$$

In the formula, the Euclidean loss function and the loss function based on the predicted number of people are accumulated to obtain the comprehensive loss function, which has been used as the loss function in the training process of this paper. Among them,  $\alpha$  represents the weight of the loss function calculation. It is proved in the Sect. 5.3 Parameter in loss function that when  $\alpha = 0.5$ , the most effective result is obtained.

Fig. 6 Algorithm flow



### 3.6 Algorithm flow

The process of crowd density estimation can be divided into two parts: training process and testing process. The training process should send original images and ground truth into the convolutional neural networks for iterative learning to obtain the optimal learning model, that is the minimum loss function value. The testing process should use the original image to test the optimal model obtained by the training process, obtain the final density map and determine the number of people by integrating and summing the density map. The process is shown in Fig. 6 and algorithm 1. The specific training process and testing process are explained in Algorithm 1.

## 4 Experiment

The experiment was conducted on the PC with 16 GB RAM and 2.5 GHz *Intel®Core™ i5-10300H* CPU, and accelerated by NVIDIA GTX 2060 graphics card. In this experiment, Adam is selected as the optimization method for network training, the network learning rate is  $1 \times 10^{-6}$ , the attenuation factor of the learning rate is 0.005, and the Relu function is used as the activation function to improve the training speed and effectively avoid the disappearance of the gradient and explosion. To verify the ability of the network in this paper to solve the problem of perspective distortion and complex background interference in

---

#### Algorithm1: Training process, Testing process

---

##### Training process

---

```

1: Input: Original image ,Ground truth
2: Output: The optimal model obtained by training
3: function:
4:     for i in epoch:
5:         Density (X,Y) = Convolution neural network(input)
6:         D(X,Y) = Density (X,Y)×8
7:         loss = Loss function(D(X,Y), GT)
8:         if Min loss:
9:             save model
10:        Else
11:            back propagation update parameter
12:        end if
13:    end for
14:    return optimal model
15: end function

```

---

##### Testing process

---

```

1: Input: Original image
2: Output: Predicted number of people
3: function:
4:     Density (X) = Convolution neural network(input)
5:     Count = sum(Density (X,Y))
6:     return count
7: end function

```

---

large-scale and complex population counting, experiments on four datasets: ShanghaiTech Part\_A, ShanghaiTech Part\_B [13], UCF\_CC\_50 [3], WorldExpo`10 [29]; the experiment content includes four parts: the first experiment

is to test on the dataset; the second experiment is the comparative experiment, to prove the advancement and superiority of this network by comparing evaluation indicators, the main comparison network include mainstream models



such as Switching-CNN[6], MCNN[13], CSRnet[14] and the latest models such as MSCANet[38] and DENet[39]; the third experiment is actual experiment, by counting the actual crowd images to test the general applicability of the network in this paper; the fourth experiment is ablation study, to prove the validity and rationality of the network we proposed, the performance of each module is tested.

#### 4.1 Datasets

In this paper, experiments are conducted on four datasets: ShanghaiTech Part\_A, ShanghaiTech Part\_B, UCF\_CC\_50, WorldExpo`10.

**ShanghaiTech dataset** Shanghai tech dataset contains 1198 labeled images, a total of 330,165 people. This dataset consists of two parts: Part A contains 482 images of highly crowded crowd, 300 images are used for training process, 182 images are used as testing process; Part B contains 719 images of relatively sparse crowd, 400 images are used for training process, 316 images are used as testing process.

**UCF\_CC\_50** UCF\_CC\_50 contains 50 images from the Internet with different viewing angles and different resolutions. The number of labeled in each image ranges from 94 to 4,543, with an average of 1,280. Because the number of pictures in this dataset is limited, and the number of annotated people in the pictures is also compared, five-fold cross-validation is used in this dataset [2].

**WorldExpo`10** The WorldExpo`10 dataset contains 5 scenes with 3980 pictures, of which 3380 pictures are used as the training dataset, and the remaining 600 pictures are used as the testing dataset. Simultaneously, the dataset provides a region of interest (ROI).

A representative image is selected from each dataset, as shown in Fig. 7. (a) is from ShanghaiTech Part\_A, this part of the image has the problem of complex background interference; (b) is from ShanghaiTech Part\_B, this part of the image has the problem of human head scale diversity caused by perspective distortion; (c) is from UCF\_CC\_50, the crowd in the image of this dataset is very crowded, and there is almost no interference and noise. (d) is from the WorldExpo`10 dataset.,the images in this dataset provide the uneven distribution crowd with the form of ROI regions.

#### 4.2 Ground truth generation

The existing datasets generally give the original image, the coordinate position of the corresponding crowd in the image and the total number of people. Geometric adaptive kernel

is used to generate the ground truth of the large-scale dense crowd. The definitions of geometric adaptive kernel  $F(X)$  is as follows:

$$H(X) = \sum_{i=1}^N \delta(X - X_i) \quad (9)$$

$$F(X) = H(X) \times G_{\sigma_i}(X), \text{ with } \sigma_i = \beta \bar{d}_i \quad (10)$$

where  $X_i$  is the coordinate position of the center of the human head,  $\delta(X - X_i)$  is the impact function,  $N$  is the number of labeled human heads,  $G_{\sigma_i}(X)$  is the geometric adaptive Gaussian kernel,  $\sigma_i$  is the reasonable estimation of geometric distortion,  $\beta$  is the fixed value,  $\bar{d}_i = \frac{1}{k} \sum_{k=1}^j d_j^i$  is the average of the sum of the distance between the  $i^{\text{th}}$  human head in the image and its adjacent human heads. In the experiment, we set  $\beta = 0.3$  and  $k = 3$ .

#### 4.3 Evaluation metrics

In this paper, the mean absolute error (MAE) and mean square error (MSE) are used as the evaluation indexes of the algorithm performance. Their expressions are as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{C}_i - C_i| \quad (11)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{C}_i - C_i)^2} \quad (12)$$

where  $N$  represents the number of test images,  $C_i$  represents the actual number of people in the  $i^{\text{th}}$  image, and  $\hat{C}_i$  represents the estimated number of people in the  $i^{\text{th}}$  image. When the values of MAE and MSE are smaller, the error between the estimated number of people and the actual number of people is smaller, indicating that the effect of the experiment is better.

#### 4.4 Experiment one: Dataset test

In ShanghaiTech part\_A, ShanghaiTech Part\_B and UCF\_CC\_50 datasets to test the network model's ability to solve the problem of perspective distortion and complex background interference in large-scale complex crowd density estimation, and compare the actual number of people with the predicted number. The test result is shown in Fig. 8, Fig. 9, Fig. 10.

As shown in Fig. 8, three representative sets of results from ShanghaiTech part\_A, including the crowd image with complex background interference, the corresponding ground

truth, density map, the actual number and predicted number, the test results show that the average error of crowd count is less than 4%, and the complex background interference problem in large-scale dense crowd is well solved; As shown in Fig. 9, three representative sets of results from ShanghaiTech part\_B, including the crowd image with perspective distortion caused by shooting angles, the corresponding ground truth, density map, and the actual number and predicted number, the test results show that the average error of crowd count is less than 3.5%, and the influence of perspective distortion on the accuracy of crowd density estimation is greatly reduced in large-scale dense crowd; As shown in Fig. 10, three representative sets of results from UCF\_CC50, including the large-scale dense crowd image, the corresponding ground truth, density map, and the actual number and predicted number, the test results show that the average error of crowd count is less than 1.5%, and the ideal result is also achieved in the aspect of super large-scale dense crowd density estimation. The test results on three datasets show that the predicted number of people is close to or even the same as the real number. Therefore, the network model proposed in this paper can better solve the problems of background interference and perspective distortion in large-scale dense crowd density estimation under complex background.

#### 4.5 Experiment two: Comparative experiments

To prove the advancement and superiority of the network model in this paper, performance tests were performed on the Shanghai Tech dataset and UCF\_CC\_50 dataset, and compared with some proposed network structures and some newly proposed network structures, using MAE and MSE as the measurement indicators. The results are shown in Table 1, the best results are shown in bold. The compared network structures are Switching-CNN [6], MCNN [13], CSRNet [14], FDCnet [16], ACSCP [40], CP-CNN [41], Cascaded-MTL [42], IG-CNN [43], SANet [44], D-ConvNet [45], DENet [39], SegCrowdNet [46], MSCANet [38], etc. The MAE and MSE of the above network structure are all from the original paper.

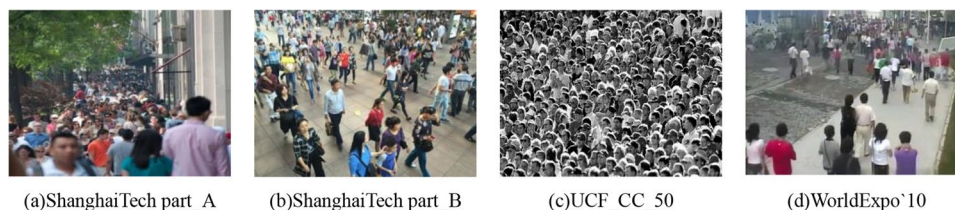
As shown in Table 1, in ShanghaiTech Part\_A, the test results of the network model proposed in this paper show that both MAE and MSE have reached the minimum, that is, the network model in this paper is better than the existing mainstream network models, compared

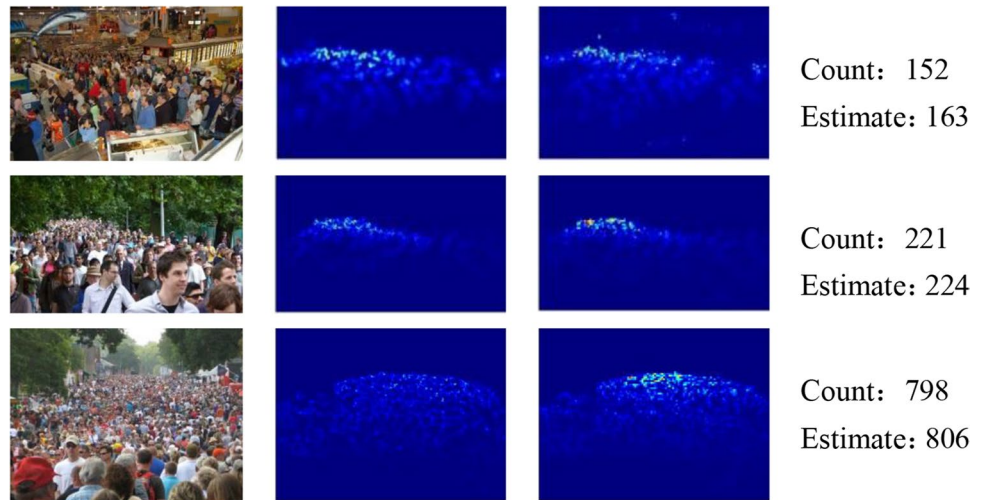
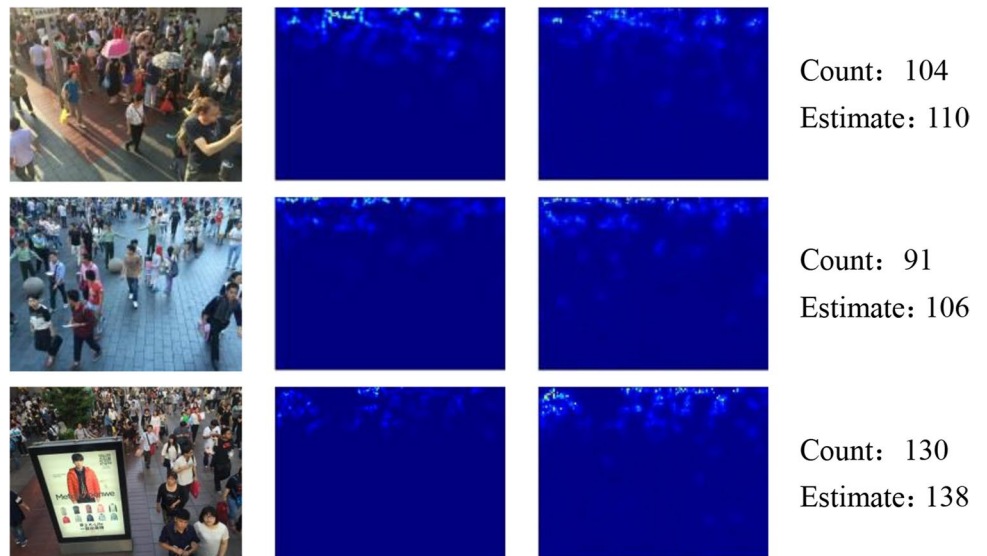
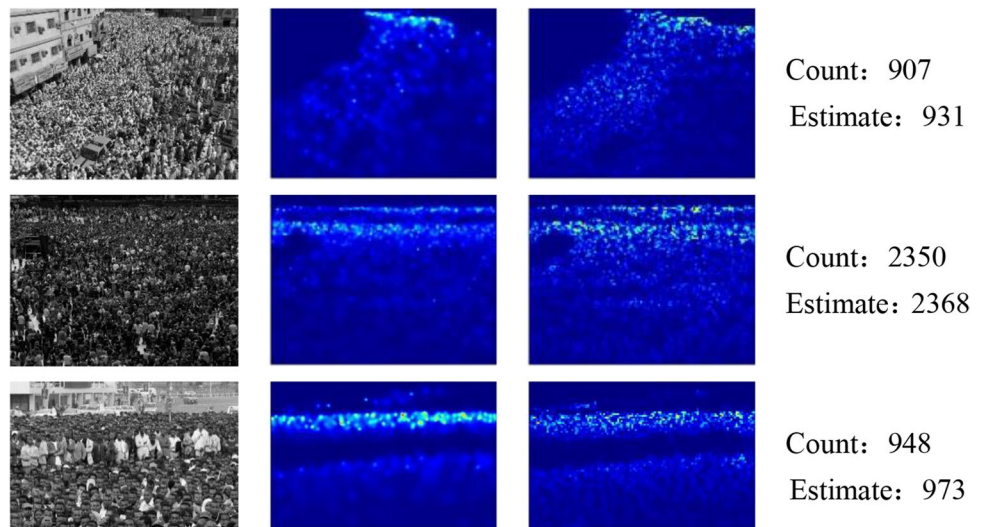
with the current network model DENet with good performance, MAE and MSE are reduced by 2.6% and 0.4% respectively; in ShanghaiTech Part\_B, the test results of the network model proposed in this paper show that the MAE of the network model proposed in this paper is 1.4% higher than SANet, and MSE is 0.4% lower than SANet; in UCF\_CC-50 dataset, the MAE of the network model proposed in this paper is 9.1% higher than SegCrowdNet and 0.8% higher than DENet, and MSE is 0.5% lower than CP-CNN. Compared with other networks, the network model proposed in this paper has different degrees of optimization in MAE and MSE. At the same time, the parameter quantity is used as a measure of the network's computational performance, the parameter quantity of the network model proposed in this paper is 11.27 million, which is lower than Switching-CNN(15.11 million), CSRNet(16.62 million), CP-CNN(68.4 Million) and other comparison networks, the amount of calculation on each dataset is also significantly reduced. As shown in Table 2, on the ShanghaiTechPart\_A dataset, the average size of the input image is  $576 \times 864$ , on the ShanghaiTechPart\_B dataset, the average size of the input image is  $768 \times 1024$ , on the UCF\_CC-50 dataset, the average size of the input image is  $656 \times 1024$ , the number of floating point operations of the network model proposed in this paper is lower than most of the comparison algorithms on the above three datasets, compared with IG-CNN, ACSCP, etc., although the calculation amount of the network model proposed in this paper is higher than the comparison algorithm, the accuracy of counting and MAE and MSE are better than comparison algorithms.

To prove that the network proposed in this paper is also suitable for the scene of uneven population distribution, we conducted a test experiment on the WorldExpo`10 dataset, and compared with some proposed network structures and some newly proposed network structures, using MAE as the measurement indicators. The results are shown in Table 3, the best results are shown in bold. The compared networks are Switching-CNN [6], MCNN [13], CSRNet [14], ACSCP [40], CP-CNN [41], IG-CNN [43], SANet [44], DENet [39], SegCrowdNet [46], The MAE of the above network structure are all from the original paper.

As shown in Table 3, in the WorldExpo`10 dataset, the Average MAE of the method proposed in this paper surpasses all the methods compared, however, not all scenarios

**Fig. 7** Representative examples from four crowd density estimation datasets



**Fig. 8** Example of ShanghaiTech part\_A dataset test**Fig. 9** Example of ShanghaiTech part\_B dataset test**Fig. 10** Example of UCF\_CC\_50 dataset test

are the best. In scenario 1, scenario 3, and scenario 5, the method proposed in this paper achieves the optimal MAE. In scenario 1, compared with the current best method SegCrowdNet, it is reduced by 0.2%; In scenario 3, compared with the current best method CSRNet, it is reduced by 0.1%; In scenario 5, compared with the current best method SegCrowdNet, it is reduced by 0.2%. The method proposed in this paper does not work very well in scenario 2 and scenario 4. The MAE in these two scenarios is 1.6% and 3% higher than the current best method, respectively, but the Average MAE of the method proposed in this paper in the five scenarios is 0.26% lower than the current best method ACSCP. On the WorldExpo`10 dataset, the number of floating point operations as shown in the last column of Table 2. The average size of the input image is  $576 \times 720$ , on this dataset, the calculation amount of the network model proposed in this paper is also lower than that of most comparison algorithms. Therefore, the method proposed in this paper

applies also to scenarios where the population is unevenly distributed.

In summary, compared with the existing network models, and the network model in this paper shows obvious advancement and superiority. In 5.4 Statistical analysis of comparative experiments, Friedman tests were used to prove and discuss the analysis.

#### 4.6 Experiment there: Actual experiments

To test the universal applicability of this network model, a real scene test is conducted. The test data come from crowd images taken on campus and urban roads. These images have complex background interference problems and perspective distortion problems. The test results are shown in Fig. 11.

Here are six sets of test results, including original pictures, density maps, and actual and predicted numbers of people. The test pictures given include crowd images with a uniform perspective distortion caused by shooting angles,

**Table 1** Comparison of network models on ShanghaiTech dataset and UCF\_CC\_50 dataset

Method	Shanghai Tech Part_A		Shanghai Tech Part_B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3	377.6	509.1
Switching-CNN	90.4	135	21.6	33.4	318.1	439.2
Cascaded-MTL	101.3	152.4	20.2	31.1	322.8	397.9
D-ConvNet	73.5	112.3	18.7	26	288.4	404.7
IG-CNN	72.5	118.2	13.6	21.1	291.4	349.4
ACSCP	75.5	102.7	17.2	27.4	291	404.6
CSRNet	68.2	115	10.6	16	266.1	397.5
MSCANet	66.5	102.1	-	-	242.8	329.8
CP-CNN	73.6	106.4	20.1	30.1	295.8	320.9
FDCnet	75.1	118.5	10.3	15.8	246.8	322.2
DENet	65.5	101.2	9.6	15.4	241.9	334.9
SegCrowdNet	68.3	104.1	12.1	19.3	<b>233.6</b>	352.6
SANet	67	104.5	<b>8.4</b>	13.6	258.4	334.9
Our method	<b>62.9</b>	<b>100.8</b>	9.8	<b>13.2</b>	242.7	<b>320.4</b>

**Table 2** The inference efficiency of Comparison network models. #Param denotes the number of parameters, while FLOPs is the number of floating point operations. The units are million (M) for #Param, giga (G) for FLOPs

Method	#Param	Part_A FLOPs( $576 \times 864$ )	Part_B FLOPs( $768 \times 1024$ )	UCF_CC_50 FLOPs( $656 \times 1024$ )	WorldExpo`10 FLOPs( $576 \times 720$ )
Switching CNN	15.11	191.24	302.16	258.11	159.35
D-ConNet	16.62	210.33	332.36	283.91	175.28
IG-CNN	4.7	59.48	93.98	80.28	49.57
ACSCP	5.1	64.54	101.99	87.12	53.78
CSRNet	16.26	205.88	325.34	277.42	171.57
CP-CNN	68.4	674.59	1065.96	910.31	562.17
FDCnet	12.98	165.84	262.06	223.85	138.19
Our method	11.27	96.37	152.28	130.08	80.31



crowd images with complex background interference problems, and large-scale dense crowd images. As shown in Fig. 11, the network model proposed in this paper can also generate high-quality density maps for test samples that have not been trained. Analyze the predicted number of crowds obtained, and compare the actual number of people with the predicted number of people. It can be seen that there are errors in the prediction within a reasonable range. Therefore, the network model proposed in this paper is universally applicable, it has strong application value for large-scale perspective distortion crowd density estimation under complex background.

#### 4.7 Experiment four: Ablation study

Since the network structure proposed in this paper is composed of multiple modules, we have conducted ablation experiments on ShanghaiTech dataset to prove the effectiveness of the IA-MFFCN structure. The experimental results are shown in Table 4.

We first performed an experiment on the feature extraction module, and the test results are shown in the first row of Table 4, it can be seen that neither MAE nor MSE in the first row is the smallest, that is, it cannot provide the best performance; then, after the feature extraction module, the inverse attention mechanism was added, and the other structures remain unchanged. It can be seen that compared to only have feature extraction module, MAE and MSE on the ShanghaiTech Part\_A dataset are reduced by 15% and 11.7% respectively, and MAE and MSE on the ShanghaiTech Part\_B dataset reduced by 1.8% and 9.8%, respectively, which fully proves the gain of the inverse attention module to the entire model; finally, using IA-MFFCN for experiments, the test results show that compared with the above two experiments, MAE and MSE on the ShanghaiTech Part\_A dataset are reduced by 2.5% and 7.8% respectively, and MAE and MSE on the ShanghaiTech Part\_B dataset are reduced by 4.8% and 2.1% respectively, which proves that each module has a certain effect on

improving network performance and verifies the validity and rationality of our proposed network module.

## 5 Discussion

Since our network is composed of multiple modules, we should verify their effectiveness to obtain the final performance. Because the ShanghaiTech dataset is the dataset with the largest number of labeled people, this paper discusses based on this dataset in the same experimental environment.

### 5.1 Study of Feature extraction module

To verify the ability of the feature extraction module of this paper to extract features of different scales, without changing other structures, the first 10 layers of VGG16 are used to replace the Feature extraction module to verify its impact on network performance. The experimental results are shown in Table 5. Among them, VGG represents using the first ten layers of VGG16 as the feature extraction network, and FEM represents the feature extraction module proposed in this paper, which is the method proposed in this paper.

It can be seen from Table 5 that using the feature extraction module proposed in this paper (the second row of the table) compared with using the first 10 layers of VGG16 as the feature extraction module (the first row of the table) has a significant reduction in MAE and MSE, namely, The feature extraction module proposed in this paper has obvious advantages in feature extraction ability and counting accuracy.

### 5.2 Study of Inverse attention module

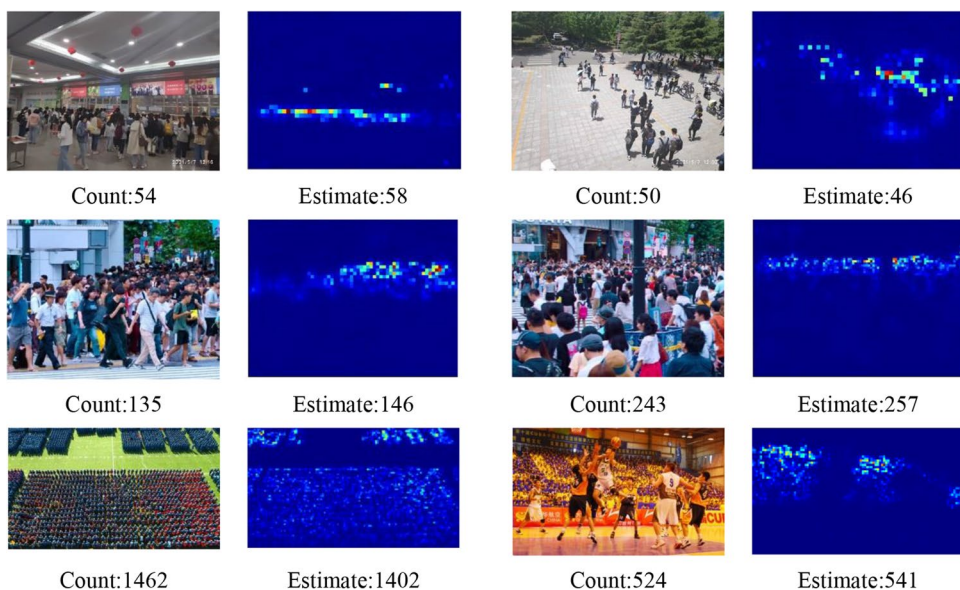
The traditional inverse attention mechanism only uses a list of networks for feature extraction, which may easily cause the loss of information, which will affect the performance

**Table 3** Comparison of network models of five scenarios on WorldExpo`10 dataset

Method	S1	S2	S3	S4	S5	Average
MCNN	3.4	20.6	12.9	13	8.1	11.6
IG-CNN	2.6	16.1	10.15	20.2	7.6	11.3
Switching-CNN	4.4	15.7	10	11	5.9	9.4
CSRNet	2.9	11.5	8.6	16.6	3.4	8.6
CP-CNN	2.9	11.5	8.6	16.6	3.4	8.6
SANet	2.8	14	10.2	12.5	3.5	8.6
DENet	2.8	<b>10.7</b>	8.6	15.2	3.5	8.2
SegCrowdNet	2.3	11.9	11.8	11.2	2.4	7.9
ACSCP	2.8	14.05	9.6	<b>8.1</b>	2.9	7.5
Our method	<b>2.1</b>	12.3	<b>8.5</b>	11.1	<b>2.2</b>	<b>7.24</b>



**Fig. 11** Results of actual experiments



**Table 4** Results of ablation study on ShanghaiTech dataset

Module			ShanghaiTech Part_A		ShanghaiTech Part_B	
Feature extraction module	MIAM	Back-end module	MAE	MSE	MAE	MSE
√			80.4	120.3	16.4	25.1
√	√		65.4	108.6	14.6	15.3
√	√	√	62.9	100.8	9.8	13.2

of the network. We have expanded the branch of the convolutional network in the inverse attention mechanism, added a pooling layer. To prove its effectiveness, a comparative experiment was conducted on the ShanghaiTech Part\_A dataset for IAB and MIAM. Both the feature extraction module and the back-end module of the experiment use the corresponding parts of the network model proposed in this paper. The results of the comparison are shown in Table 6.

The results show that IAB plays a certain role in solving the problem of complex background interference in crowd density estimation, but its feature extraction ability is limited, MIAM has a stronger ability to solve the problem of complex background interference in crowd density estimation. It is tested on the ShanghaiTech Part\_A dataset, the results show that the introduction of IAB reduces MAE by 5% and MSE by 13.6%; compared with the network

introduced with IAB, the network structure with the introduction of MIAM reduces MAE by 0.3% and MSE by 0.6%. The introduction of MIAM improves the network's ability to reduce background interference and improves the accuracy of the network's crowd density estimation.

### 5.3 Parameter in loss function

In the training process of the network, the use of different loss functions will have a direct impact on the performance of the model. This article uses the Euclidean loss function and the loss function based on the predicted number of people as the loss function of this article, and adjust the proportion of the two by calculating the weight of the loss to get the best results, experiments on the ShanghaiTech dataset. The experimental results are shown in Fig. 12.

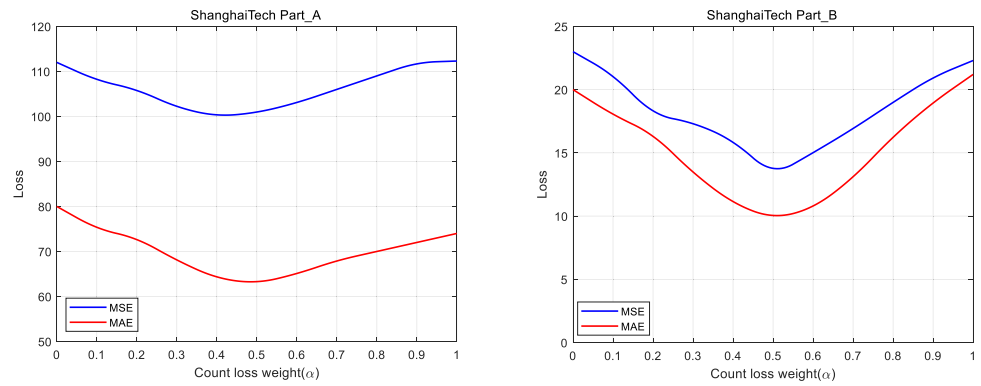
**Table 5** Comparison results of feature extraction modules

Method	ShanghaiTech Part_A		ShanghaiTech Part_B	
	MAE	MSE	MAE	MSE
VGG	72.4	108.3	11.7	16.8
FEM	62.9	100.8	9.8	13.2

**Table 6** Discussion results of ShanghaiTech Part\_A dataset

Method	MAE	MSE
Feature extraction module + Back-end module	68.2	115
Feature extraction module + IAB + Back-end module	63.2	101.4
Feature extraction module + MIAM + Back-end module	62.9	100.8

**Fig. 12** MAE and MSE in ShanghaiTech dataset under different counting weights



As shown in Fig. 12, it shows the changes of MAE and MSE in the ShanghaiTech dataset as  $\alpha$  changes. With the change of  $\alpha$ , in the two datasets, the values of MAE and MSE show a trend of first decreasing and then increasing. When  $\alpha = 0.5$ , that is, when the two parts of the loss function occupy the same proportion, achieved the optimal value.

**5.4 Statistical analysis of comparative experiments**

To prove the superiority of the proposed method compared with the comparison algorithm, this paper uses the Friedman test and Nemenyi test to perform a full statistical analysis on the MSE of 13 comparison algorithms on the ShanghaiTech Part\_A, Part\_B and UCF\_CC\_50 data sets. The ordinal statistical results of algorithm performance are shown in Table 7.

The results show that the average MSE order value of the method proposed in this paper is first in the comparison algorithm, which shows that the algorithm proposed in this paper improves the accuracy of crowd counting to a certain

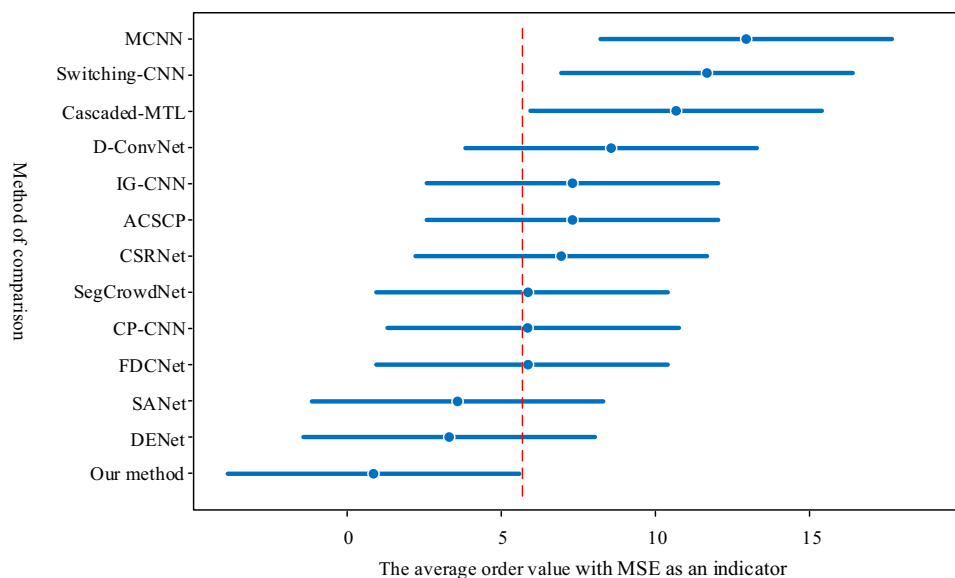
extent. Taking MSE as the measurement index, the results of the Friedman test and Nemenyi test for each algorithm are shown in Fig. 13.

The vertical axis in the figure shows the comparison algorithms, and the horizontal axis shows the average sequence value of each algorithm in the ShanghaiTech Part\_A, Part\_B and UCF\_CC\_50 data sets according to the MSE value ranking. For each algorithm, a dot is used to represent the average sequence value, and the horizontal line segment centered on the dot represents the size of the critical value range. When the critical value ranges of the two comparison algorithms do not intersect, it means that the performance of the former is significantly better than the latter of the two comparison algorithms. As shown in Fig. 13, compared with the other 12 algorithms on 3 data-sets, the method proposed in this paper shows that it has no intersection with the horizontal lines of MCNN, switching-CNN, and Cascaded-MTL, that is, it is significantly better than the above three algorithms. As can be seen, compared with other algorithms, the average sequence

**Table 7** Ordinal table of algorithm performance with MSE as an indicator

Method	ShanghaiTech Part_A	ShanghaiTech Part_B	UCF_CC_50	Average sequence
Ours	1	1	1	1
DENet	2	3	5	3.33
SANet	5	2	4	3.67
FDCnet	10	4	3	5.67
SegCrowdNet	4	6	7	5.67
CP-CNN	6	10	2	6
CSRNet	8	5	8	7
ACSCP	3	9	10	7.33
IG-CNN	9	7	6	7.33
D-ConvNet	7	8	11	8.67
Cascaded-MTL	12	11	9	10.67
Switching-CNN	11	12	12	11.67
MCNN	13	13	13	13

**Fig. 13** Friedman test chart with MSE as a measurement indicator



ranking of the MSE value of the method proposed in this paper is 1, while the average sequence ranking of the MSE value of DENet is 3.33, which has the second best performance. That is to say, the comprehensive performance of the three datasets of ShanghaiTech Part\_A, Part\_B and UCF\_CC\_50 proposed in this paper has a certain degree of advantage, which reduces the value of MSE and improves the accuracy of crowd counting. In summary, compared with the existing network models, and the network model in this paper shows obvious advancement and superiority.

## 6 Conclusion

To solve the problem of difficulty in counting caused by perspective distortion and background interference, in this paper, we propose a multi-scale feature fusion network based on the inverse attention mechanism (IA-MFFCN) to estimate the number of crowd with large-scale perspective distortion in complex backgrounds. IA-MFFCN uses a single-column convolution kernel multi-scale feature fusion network to extract features of different scales and perform multi-scale feature fusion to solve the problem of perspective distortion, uses a multi-column inverse attention module to denoise the interference background, reduce the impact of background noise on counting accuracy, and solve the background interference problem, use a new type of comprehensive loss function to enhance the sensitivity of the counting network, finally, the quality of the density map is improved by dilated convolution. The experimental results on the ShanghaiTech dataset, UCF\_CC\_50 dataset and WorldExpo'10 dataset show that the method in this paper can effectively extract multi-scale crowd features, reduce the interference of complex backgrounds and improve the accuracy of crowd counting. In future work, we

will further optimize our method, such as preprocessing the image, using other methods to generate density maps, etc., to improve its accuracy and robustness. Simultaneously, we will further transfer our work to more crowd-counting scenes, especially scenes where crowds move faster, such as train stations and airports. Additionally, the work of crowd counting can be transferred to other counting fields, such as highway vehicle density counting, port ship density counting, etc.

**Acknowledgements** The authors are grateful for the collaborative funding support from the Humanity and Social Science Foundation of Ministry of Education, China (21YJAZH077).

## References

- Xu M, Li C, Lv P, Lin N, Hou R, Zhou B (2017) An efficient method of crowd aggregation computation in public areas. *IEEE Trans Circuits Syst Video Technol* 28(10):2814–2825
- Idrees H, Soomro K, Shah M (2015) Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE Trans Pattern Anal Mach Intell* 37(10):1986–1998
- Idrees H, Saleemi I, Seibert C, Shah M (2013). Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* 2547–2554
- Hu C, Wang Y, Gu J (2020) Cross-domain intelligent fault classification of bearings based on tensor-aligned invariant subspace learning and two-dimensional convolutional neural networks. *Knowledge-Based Systems* 209:106214
- Hu C, He S, Wang Y (2021) A classification method to detect faults in a rotating machinery based on kernelled support tensor machine and multilinear principal component analysis. *Appl Intell* 51(4):2609–2621
- Babu Sam D, Surya S, Venkatesh Babu R (2017). Switching convolutional neural network for crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* 5744–5752

7. Onoro-Rubio D, López-Sastre RJ (2016). Towards perspective-free object counting with deep learning. In European conference on computer vision. Springer, Cham, 615–629
8. Wang L, Yin B, Guo A, Ma H, Cao J (2018) Skip-connection convolutional neural network for still image crowd density estimation. *Appl Intell* 48(10):3360–3371
9. Jiang M, Lin J, Wang ZJ (2021) A smartly simple way for joint crowd counting and localization. *Neurocomputing* 459:35–43
10. Xia Y, He Y, Peng S, Hao X, Yang Q, Yin B (2021) EDENet: Elaborate density estimation network for crowd counting. *Neurocomputing* 459:108–121
11. Wang W, Liu Q, Wang W (2021). Pyramid-dilated deep convolutional neural network for crowd counting. *Appl Intell* 1–13
12. Amirgholipour S, Jia W, Liu L, Fan X, Wang D, He X (2021) PDANet: Pyramid density-aware attention based network for accurate crowd counting. *Neurocomputing* 451:215–230
13. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016). Single-image crowd density estimation via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 589–597
14. 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 (CVPR) 1091–1100
15. Nie P, Fan C, Zou L, Chen L, Li X (2020) crowd density estimation Guided by Attention Network. *Information* 11(12):567
16. Zhang Y, Li G, Lei J, He J (2019) FDCNet: Frontend-backend fusion dilated network through channel-attention mechanism. *Appl Sci* 9(17):3466
17. Li T, Chang H, Wang M, Ni B, Hong R, Yan S (2014) Crowded scene analysis: A survey. *IEEE Trans Circuits Syst Video Technol* 25(3):367–386
18. Wang L, Yin B, Tang X, Li Y (2019) Removing background interference for crowd density estimation via de-background detail convolutional network. *Neurocomputing* 332:360–371
19. Lin SF, Chen JY, Chao HX (2001) Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 31(6):645–654
20. 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
21. 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 (ICPR), 1–4. IEEE
22. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2009) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32(9):1627–1645
23. Dollar P, Wojek C, Schiele B, Perona P (2011) Pedestrian detection: An evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell* 34(4):743–761
24. An S, Liu W, Venkatesh S (2007). Face recognition using kernel ridge regression. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–7
25. Chan AB, Vasconcelos N (2009). Bayesian poisson regression for crowd density estimation. In 2009 IEEE 12th international conference on computer vision (ICCV) 545–551
26. Pham VQ, Kozakaya T, Yamaguchi O, Okada R (2015). Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3253–3261
27. Lempitsky V, Zisserman A (2010) Learning to count objects in images. *Adv Neural Inf Process Syst* 23:1324–1332
28. Chan AB, Vasconcelos N (2011) Counting people with low-level features and Bayesian regression. *IEEE Trans Image Process* 21(4):2160–2177
29. Zhang C, Li H, Wang X, Yang X (2015). Cross-scene crowd density estimation via deep convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 833–841
30. Sindagi VA, Patel VM (2017) Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd density estimation. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) 1–6. IEEE
31. Liu W, Salzmann M, Fua P (2019) Context-aware crowd density estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 5099–5108
32. Wang Q, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd density estimation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 8198–8207
33. Zhu F, Yan H, Chen X, Li T, Zhang Z (2021) A multi-scale and multi-level feature aggregation network for crowd counting. *Neurocomputing* 423:46–56
34. Liu YB, Jia RS, Liu QM, Zhang XL, Sun HM (2021) Crowd counting method based on the self-attention residual network. *Appl Intell* 51(1):427–440
35. Gu L, Pang C, Zheng Y, Lyu C, Lyu L (2021) Context-aware pyramid attention network for crowd counting. *Appl Intell*.1–17
36. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, ..., Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (2048–2057). PMLR
37. Sindagi VA, Patel VM (2019). Inverse attention guided deep crowd density estimation network. In 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS) 1–8. IEEE
38. Zhang Y, Zhao H, Duan Z, Huang L, Deng J, Zhang Q (2021) Congested crowd density estimation via Adaptive Multi-Scale Context Learning. *Sensors* 21(11):3777
39. Liu L, Jiang J, Jia W, Amirgholipour S, Wang Y, Zeibots M, He X (2020) Denet: A universal network for counting crowd with varying densities and scales. *IEEE Trans Multimedia* 23:1060–1068
40. Shen Z, Xu Y, Ni B, Wang M, Hu J, Yang X (2018) crowd density estimation via adversarial cross-scale consistency pursuit. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 5245–5254
41. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid cnns. In Proceedings of the IEEE international conference on computer vision (ICCV) 1861–1870
42. Sindagi VA, Patel VM (2017) Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd density estimation. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) 1–6
43. Sam DB, Sajjan NN, Babu RV, Srinivasan M (2018) Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 3618–3626
44. Cao X, Wang Z, Zhao Y, Su F (2018). Scale aggregation network for accurate and efficient crowd density estimation. In Proceedings of the European Conference on Computer Vision (ECCV) 734–750
45. Shi Z, Zhang L, Liu Y, Cao X, Ye Y, Cheng MM, Zheng G (2018) crowd density estimation with deep negative correlation learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 5382–5390
46. Chen J, Wang Z (2021) Crowd counting with segmentation attention convolutional neural network. *IET Image Proc* 15(6):1221–1231

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