IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2025, vol. 3, no. 3, pp: 1–11

Crowd Density Estimation Based on Multi-scale Feature Fusion and Information Enhancement

Lina Zou[1]

Artificial Intelligence College, Shenyang Normal University
Shenyang, 110034 China
Received May. 28, 2025; Revised and Accepted July. 4, 2025

Abstract. Aiming at the problems such as diverse target scales and large-scale changes in crowds in dense crowd scenarios, a crowd density estimation method based on multi-scale feature fusion and information enhancement is proposed. Firstly, considering that small-scale targets account for a relatively large proportion in the image, based on the VGG-16 network, the dilated convolution module is introduced to mine the detailed information of the image. Secondly, in order to make full use of the multi-scale information of the target, a new context-aware module is constructed to extract the contrast features between different scales. Finally, considering the characteristic of continuous changes in the target scale, a multi-scale feature aggregation module is designed to enhance the sampling range of dense scales and multi-scale information interaction, thereby improving the network performance. Experiments on public datasets show that the proposed method in this paper can effectively estimate the population density compared with other advanced methods.

Keywords: Crowd density estimation, Multi-scale feature fusion, Information enhancement, VGG-16 network.

1. Introduction

The main task of crowd density estimation is to estimate the total number of people in a scene. It has been widely applied in fields such as public security management, urban spatial planning, and traffic dispatching, and has received considerable attention from researchers both at home and abroad. With the continuous growth of urban population, various gatherings of people are frequently held [1,2]. For instance, large crowds often gather at tourist attractions, large stadiums, and popular business districts. The demand for crowd counting is increasing day by day. However, in actual scenarios, due to issues such as diverse target scales, continuous changes in the scale of the same image, and significant density differences among images, the task of crowd counting still faces considerable challenges [3].

At present, traditional research schemes for population density estimation can be roughly divided into two categories, namely, methods based on pedestrian detection and methods based on population regression. The limitation of the pedestrian detection method lies in that when there is occlusion among the crowd, most pedestrians cannot be accurately detected, resulting in the counting result being much smaller than the actual number of people [4]. Although the regression method based on the number of people can solve the occlusion problem, a single number of people result cannot reflect the distribution of the crowd and spatial information in the scene [5].

In recent years, due to the powerful feature extraction ability of convolutional neural networks (CNN), they have been widely applied in the field of crowd counting and achieved remarkable results [6]. Li et al. [7] first proposed a crowd counting model in combination with CNN, but they did not consider the impact of target scale changes on model performance, resulting in low counting accuracy. Based on this, Ranasinghe et al. [8] proposed a multi-column convolutional neural network (MCNN), which used three branches with different convolutional kernels to construct a network model to capture multi-scale features under different receptive fields.

The proposal of MCNN laid the foundation for multi-branch population counting research, but its model structure has redundancy and the computational cost is high. Therefore, Gupta et al. [9] proposed CSRNet (congested scene recognition network) by combining the first 10 layers of VGG-16 and dilated convolution, which could simplify the structure while better aggregating multi-scale features in crowded scenes. Considering that dilated convolution had the advantage of expanding the receptive field without increasing the computational load, Seo et al. [10] constructed a multi-scale feature extraction module to improve the accuracy of crowd counting based on dilated convolution. Wang et al. [11] proposed a scale pyramid network (SPN), which adopted a parallel single-column structure and constituted a scale pyramid module through dilated convolution to extract deep multi-scale information. Alotaibi et al. [12] introduced a multi-level bottom-up and top-down fusion network from the perspective of feature fusion. By interacting shallow and deep information in a bidirectional manner across different scales, the effectiveness of multi-scale fusion was enhanced. Patidar et al. [13] proposed a multi-scale generative adversarial network, which utilized fused features from different levels to detect large-scale changing populations

and estimated the population density through an adversarial training model. Jaiswal et al. [14] utilized the trellis encoder-decoder network (TEDNet) to hierarchically aggregate features and improve the expression of scale-varying targets. Kamra et al. [15] to alleviate the problem of insufficient generalization under different population densities, constructed multiple pre-trained sub-networks in different density scenarios to mine general density information. In addition, references [16-19] also utilized contextual information to optimize counting tasks, thereby enhancing the adaptability and accuracy of the model in complex scenarios.

Although the above-mentioned method can obtain the feature information of the crowd, it only processes the input image through simple feature extraction and ignores the characteristic of continuous change in the target scale. Therefore, how to utilize network models to extract the characteristics of people with continuously changing scales, reduce the loss of spatial detail information, and effectively integrate multi-level scale features remains an urgent problem to be solved. Therefore, how to utilize network models to extract the characteristics of people with continuously changing scales, reduce the loss of spatial detail information, and effectively integrate multi-level scale features remains an urgent problem to be solved. To this end, this paper proposes a dense crowd counting network based on multi-scale perception. The network structure is mainly composed of the dilated convolution module (DCM), the context-aware module (CAM), and the multi-scale feature aggregation module (MSAM). Specifically, the primary features extracted by VGG-16 are respectively processed through the DCM and CAM modules to obtain rich fine-grained and contextual information. Then, the MSAM module is utilized to extract multi-scale features and achieve effective aggregation. Finally, the standard convolution is applied to obtain the final predicted density map. Experimental verification is conducted on the datasets of ShangHai Tech7, UCF-CC-5021,UCF-QNRF and NWPU. The results show that the proposed method in this paper has better counting performance.

2. Proposed Crowd Density Estimation Model

The structure of the dense crowd counting network based on multi-scale perception proposed in this paper is shown in Figure 1. In this figure, Conv3-256-2 indicates that in the convolution operation, the convolution kernel is 3, the number of channels is 256, and the dilated rate is 2. $3 \times Conv$ indicates that the convolution of this layer is performed three times.

Based on the experience of crowd counting, it can be known that the VGG-16 network mainly uses standard convolution of size 3×3, with a simple and flexible structure. It is often used as a primary feature extractor. Therefore, in this paper, the first 10 layers of the VGG-16 network are used as the backbone network. The input image I is extracted through the backbone network to obtain the feature f_v, as shown in equation (1).

$$f_v = F_{vgg}(I). \tag{1}$$

Where F_{vgg} is feature extraction function.

Considering that there are a large number of small-scale targets in dense crowd images, a dilated convolution module with a low void rate is designed after the backbone network to mine the detailed information of the image. The specific structure is shown in the DCM in Figure 1. In addition, to enhance the network's perception ability over large-scale ranges, this paper designs a context-aware module. It mainly pools the input features at different scales to learn context information and fuses it with the output features of the dilated convolution module to obtain rich multi-scale information. On this basis, a multi-scale feature aggregation module is also proposed to further aggregate multi-level features to cope with the continuous changes in scale. Finally, through standard convolution, high-quality density maps are generated, thereby achieving crowd image counting.

2.1. Context-aware Module

In the PSPNet (pyramid scene parsing network) model proposed by Zhao et al. [20], the pyramid pooling module (PPM) designed by using pooling branches of different scales can effectively aggregate context information, but there is a lack of information interaction among features of each scale. Therefore, this paper proposes a context-aware module based on PPM, and its structure is shown in CAM in Figure 1.

Firstly, the input features are divided into four different groups from coarse to fine according to scale, and pooling. 1×1 convolution operations are performed on each group respectively. The sizes of the pooling kernels in each group are 1×1, 2×2, 3×3, and 6×6 respectively. Then, bilinear interpolation is used for up-sampling to obtain scale-aware features $f_{c,j}$ of the same size as the input, as shown in equation (2).

$$f_{c,j} = U_b(F_0(P_{ave}(f_v, j), \theta_j)). \tag{2}$$

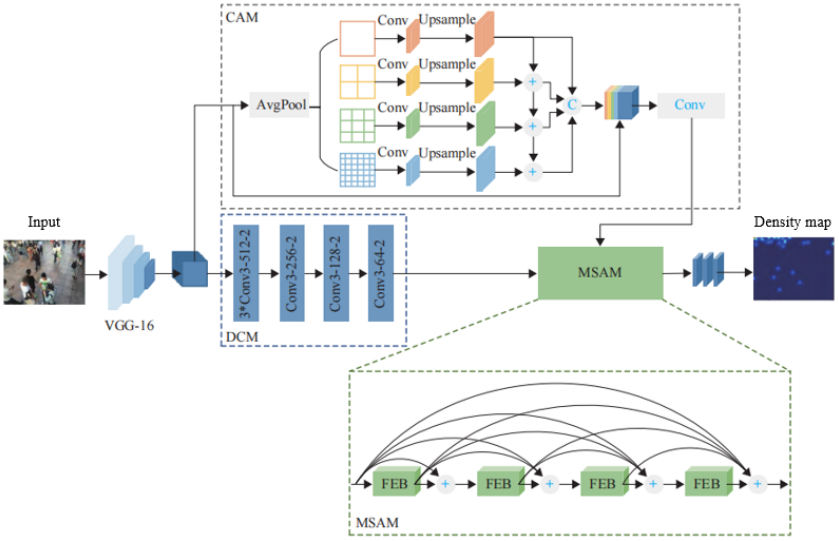

Fig. 1. Proposed network

In the formula, j represents the scale, and $P_{ave}(\cdot)$ is the average pooling function. $F_0(\cdot)$ is a convolution operation of size 1×1. $U_b(\cdot)$ is a bilinear interpolation function. θ_j is the network parameter.

Secondly, in order to fully utilize the context information of different scales in the crowd images, CAM adopts a top-down and branch-by-branch addition approach to aggregate the contrast features of different branches. Then, through the feature concatenation operation, the output features of the four branches are concatenated and cross-channel fused with the original features. Finally, the fused features are input into the subsequent convolutional layers to output the final result. Introducing the CAM structure behind the backbone network can capture feature information of different scales in a hierarchical manner from coarse to fine, thereby obtaining rich multi-scale context features.

2.2. Multi-scale Feature Aggregation Module

Inspired by reference [21], introducing semantic information into shallow features and embedding spatial information into deep features can effectively integrate features of different scales. Based on this, this paper proposes a multi-scale feature aggregation module composed of FEB.

FEB is the core component of MSAM, and its structure is shown in Figure 2. Conv3-1 indicates that the convolution kernel is 3 and the void ratio is 1, and so on. Firstly, to reduce the computational complexity, the input features are evenly divided into four feature subsets after undergoing 1×1 convolution, which are x_1 to x_4 respectively. Here, the number of channels for each feature subset is reduced to 1/4 of the input features. Except for x_1, each of the other $x_k(k = 2, 3, 4)$ has a set of corresponding dilated convolution with dilation rates of $(1, 2, 3)$, $(3, 4, 5)$ and $(5, 6, 7)$ respectively. This setting method can reduce the grid effect of dilated convolution and improve information continuity. At the same time, it connects with other subsets in the form of residuals. For the $k - th$ subset, its corresponding output is:

$$f_{e,k} = \begin{cases} x_k & k = 1 \\ C_{k,3}(\sum_{n=2}^{k} C_{n,2}(\sum_{l=2}^{n} C_{l,1}(x_k))) & k = 2, 3, 4 \end{cases} \tag{3}$$

Where $C_{k,3}(\cdot)$ is the third dilated convolution operation in the $k - th$ row, and other similar structures follow this pattern.

By fusing $f_{e,k}$ and refining the feature map, the output feature FEB of the f_{feb} structure is obtained as:

$$f_{feb} = f_{in} \oplus (F_1(\sum_{k=1}^{4} f_{e,k}) \otimes \varphi(F_2(F_1(\sum_{k=1}^{4} f_{e,k})))). \tag{4}$$

Where, f_{in} is the input of the FEB structure. Both $F_1(\cdot)$ and $F_2(\cdot)$ are standard convolution. $\varphi(\cdot)$ is the sigmoid activation function, where is the addition of each element. $\oplus$ is the multiplication of each element.

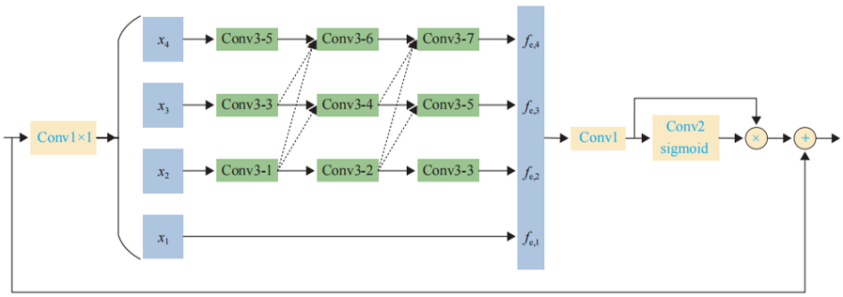

Fig. 2. Structure of feature enhancement block

In FEB, the loss of pixel information is effectively reduced by setting the void rate coprime between dilated convolution layers and closely connecting each extended layer with other layers. Meanwhile, dilated convolution with a void rate of 3 and 5 is used twice each in the network to avoid jumps in the size of the receptive field. In summary, FEB enriches the scale diversity by enhancing the continuity of the receptive field range, which is more conducive to extracting population features with continuous scale changes.

However, due to the lack of correlation between FEBs, this paper designs an MSAM structure in combination with the dense residual connection strategy [22]. By integrating the multi-scale features of different network layers, cross-layer connections between networks of different depths are achieved. Among them, integrating the relatively shallow detail information into the deep layer makes the information of the subsequent layers more abundant. Meanwhile, the reuse of features can reduce the information loss caused by the deepening of the network and improve the model performance.

2.3. Structured Density Map

Since training the network model requires estimating the crowd density map from the input crowd images, we generate the density map needed for training based on an adaptive Gaussian kernel proposed in reference [23].

Suppose there is a head at the pixel x_i position, which can be represented by the function $\delta(x - x_i)$, then an image containing the head position markers of N people can be represented as:

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i). \tag{5}$$

To convert it into a continuous density function, here a Gaussian kernel $G_\sigma(x)$ is convolution with this function to generate a continuous density map $F(x) = H(x) * G_\sigma(x)$. However, the distortion between the ground plane and the image plane caused by the shooting Angle of the image leads to the head size of each person in the image being different, thus making it impossible to determine the parameters of the Gaussian kernel σ. According to the idea in reference [24], in a crowded scene, the parameter σ of each person are adaptively determined by the position of each person's head and the average distance between the heads of adjacent people. The following presents the scheme for determining the parameter σ of the Gaussian kernel.

For each head x_i in a given crowd image, if the distance between it and its k adjacent heads is defined as $d_1^i, d_2^i, \cdots, d_k^i$, then the average distance from the k adjacent heads to head x_i can be defined as:

$$\bar{d}_i = \frac{1}{k} \sum_{j=1}^{k} d_j^k. \tag{6}$$

Here, the Gaussian kernel parameter σ_i of head x_i is determined by the average distance $\bar{d}_i$. Define $\sigma_i = \beta d_i$, $\beta = 0.3$, $k = 4$. Then the density graph F is ultimately defined as:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) G_{\sigma_i}(x). \tag{7}$$

Here, N represents the total number of heads in the image. x_i represents the coordinates of a person's head in an image. Based on this scheme, it can be understood that geometrically adaptive Gaussian kernel generation density maps are generally suitable for crowded crowd scenarios, while sparse crowds usually adopt fixed Gaussian kernels to generate density maps. For generating density maps with a fixed Gaussian kernel, the principle is similar

to that of the geometrically adaptive Gaussian kernel. The only difference is that each σ_i is fixed, and different population datasets have different σ_i.

2.4. Loss Function

In this paper, the model training first adopts Euclidean loss to calculate the difference between the estimated density map and the true density map, which is defined as follows:

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N}||I(X_i,\theta) - I_i^{gt}||^2. \tag{8}$$

Where θ represents the parameter model. $I(X_i,\theta)$ represents the output model. The X_i and I_i^{gt} respectively represent the $i-th$ original input image and the true density map in the training set.

Considering the performance of the model in sparse scenarios, this paper introduces a relative crowd loss to improve the performance of the model in sparse scenarios. The loss is defined as follows:

$$L_D(\theta) = \frac{1}{N}||\frac{I(X_i,\theta) - I_i^{gt}}{I_i^{gt}+1}||^2. \tag{9}$$

Where, the denominator $I_i^{gt}+1$ is to prevent the denominator from being zero.

Therefore, the total loss function of the model can be defined as:

$$L_{loss} = L(\theta) + \alpha \times L_D(\theta). \tag{10}$$

Here, α represents the proportion of the relative population loss in the total loss function. In this paper, we take $\alpha = 0.1$.

3. Experiment and Result Analysis

3.1. Training Details

The method code in this paper is implemented based on the Pytorch framework and experiments are conducted under the configuration of the Windows10 operating system and NVIDIA GeForce RTX 3080GPU. In addition, the model training uses the Adam optimizer, with the learning rate set to 1×10^{-4} and the momentum set to 0.9. Each batch of samples contains 10 images. To ensure the model is fully trained, random flipping and cropping operations are performed at different positions on the images to enhance the model's robustness.

3.2. Real Density Map Generation

The adaptive Gaussian kernel method is adopted to generate the true density map (D_{GT}), as shown in equation (11).

$$D_{GT} = \sum_{m=1}^{M}\delta(p-p_m)G_{\sigma_m}(x). \tag{11}$$

$$\sigma_m = \beta d_m. \tag{12}$$

Where M represents the total number of head marker points in the image. p is the image coordinate. p_m is the coordinate of the $m-th$ head marker point. $\delta(p-p_m)$ is the impact function. G_{σ_m} is a Gaussian kernel filter. σ_m represents the size of the Gaussian kernel. β is a hyperparameter with a value of 0.3. d_m is the average distance between p_m and three adjacent targets.

3.3. Evaluation Index

Mean absolute error (MAE) and root mean square error (RMSE) are commonly used evaluation criteria in crowd counting, and their definitions are shown in equations (13) and (14).

$$MAE = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} |y_{GT,i} - y_i|. \tag{13}$$

$$RMSE = [\frac{1}{N_{te}} \sum_{i=1}^{N_{te}} (y_{GT,i} - y_i)^2]^{0.5}. \tag{14}$$

Where N_{te} represents the number of test images. $y_{GT,i}$ and y_i represent the actual and predicted numbers of people in the $i-th$ test image, respectively.

3.4. Datasets

The Shanghai Tech dataset consists of two parts: Part-A and Part-B. Part-A consists of 482 dense images randomly collected from the Internet, and Part-B consists of 716 sparse images taken on the bustling streets of Shanghai.

UCF-CC-50 is the first dataset of dense crowd images [25]. This dataset contains 50 grayscale images of different resolutions, with high density and multiple complex scenes, which is very challenging. According to reference [26], the five-fold cross-validation method is used in the experiment to verify the model performance.

The UCF-QNRF dataset was proposed by Idrees et al. [27] and consisted of 1535 high-resolution dense images. The scene, Angle and light variations of this dataset are rich and diverse, and the distribution is chaotic, making the challenge very difficult.

NWPU is a large dataset publicly released by Northwestern Polytechnical University in 2020 [28], consisting of 5109 high-resolution images, including 351 negative samples, covering a variety of complex scenarios. It is currently the largest and most challenging dataset for crowd counting.

3.5. Experimental Results and Analysis

Training is conducted on four datasets and compared with advanced existing methods. The results are shown in Tables 1-5. The bold values indicate the optimal values.

Table 1. Comparison results with different methods on Shanghai Tech Part-A

Model	MAE	RMSE
MCNN [29]	110.3	173.3
CSRNet [30]	68.3	115.1
PDD-CNN [31]	64.8	99.2
TEDNet [32]	64.3	109.2
KDMG [33]	63.9	99.3
BL [34]	62.9	101.9
CAN [35]	62.4	100.1
MCANet [36]	60.2	100.3
SC2Net [37]	**59.0**	97.8
Proposed	62.6	**95.8**

From the above results, it can be seen that the proposed method in this paper has strong competitiveness on all four datasets. In the model performance comparison of the Shanghai Tech dataset, compared with SC2Net, the MAE of proposed method is slightly lost. This is because there are more background interferences in this dataset. To some extent, it affects the accuracy of model counting. However, RMSE decreases by 2.0% on Part-A and 3.5% on Part-B, indicating that proposed method has good stability. In the few-shot dataset UCF-CC-50, compared with MCANet, the MAE of proposed method decreases by 13.7% and the RMSE decreases by 13.6%. Meanwhile, for the UCF-QNRF dataset with rich scenarios, compared with BL, the MAE of proposed method decreases by 1.1% and the RMSE decreases by 4.3%. Compared with other comparison models, it shows good counting performance. This is because in this paper, on the basis of aggregating context and multi-scale information, a densely connected multi-scale feature aggregation module is constructed, thereby reducing the influence of continuous scale changes.

Table 2. Comparison results with different methods on Shanghai Tech Part-B

Model	MAE	RMSE
MCNN	26.5	41.4
CSRNet	10.7	16.1
PDD-CNN	8.9	14.4
TEDNet	8.3	12.9
KDMG	7.9	12.8
BL	7.8	12.8
CAN	7.9	12.3
MCANet	**6.9**	**11.1**
SC2Net	7.0	11.5
Proposed	7.0	**11.1**

Table 3. Comparison results with different methods on UCF-CC-50

Model	MAE	RMSE
MCNN	377.7	509.2
CSRNet	266.2	397.6
PDD-CNN	205.5	311.8
TEDNet	249.5	354.6
KDMG	238.7	346.1
BL	229.4	308.3
CAN	212.3	243.8
MCANet	181.4	258.7
SC2Net	209.5	286.4
Proposed	**156.6**	**223.4**

Table 4. Comparison results with different methods on UCF-QNRF

Model	MAE	RMSE
MCNN	277.1	426.1
CSRNet	120.4	208.6
PDD-CNN	115.4	190.3
TEDNet	113.1	188.1
KDMG	99.6	173.1
BL	88.8	154.9
CAN	107.1	183.1
MCANet	100.9	186.0
SC2Net	98.6	174.6
Proposed	**87.8**	**148.3**

Table 5. Comparison results with different methods on NWPU

Model	MAE	RMSE
MCNN	218.6	700.7
CSRNet	104.9	433.5
PDD-CNN	103.2	430.7
TEDNet	99.8	421.9
KDMG	100.6	415.6
BL	93.7	471.4
CAN	93.6	489.9
MCANet	91.5	395.7
SC2Net	89.8	348.9
Proposed	**82.0**	**300.4**

The experimental results on the UCF-CC-50 and UCF-QNRF datasets show that proposed method can achieve better accuracy in dense scenarios. Furthermore, in the NWPU dataset with a wide range of variations in the number of people, proposed method achieves the best MAE and RMSE among the comparison models. Although the addition of negative samples increases the training difficulty, it helps improve the generalization ability of the model. Moreover, experiments have fully demonstrated that proposed method has better robustness.

To further verify the prediction effect of the proposed model in this paper, Figure 3 shows some visualization prediction results of the proposed method on different datasets. As can be seen from Figure 3, the predicted density map generated by the proposed method is closer to the real density map and achieves good counting results on all four datasets, indicating that the proposed method has a good multi-scale feature extraction ability.

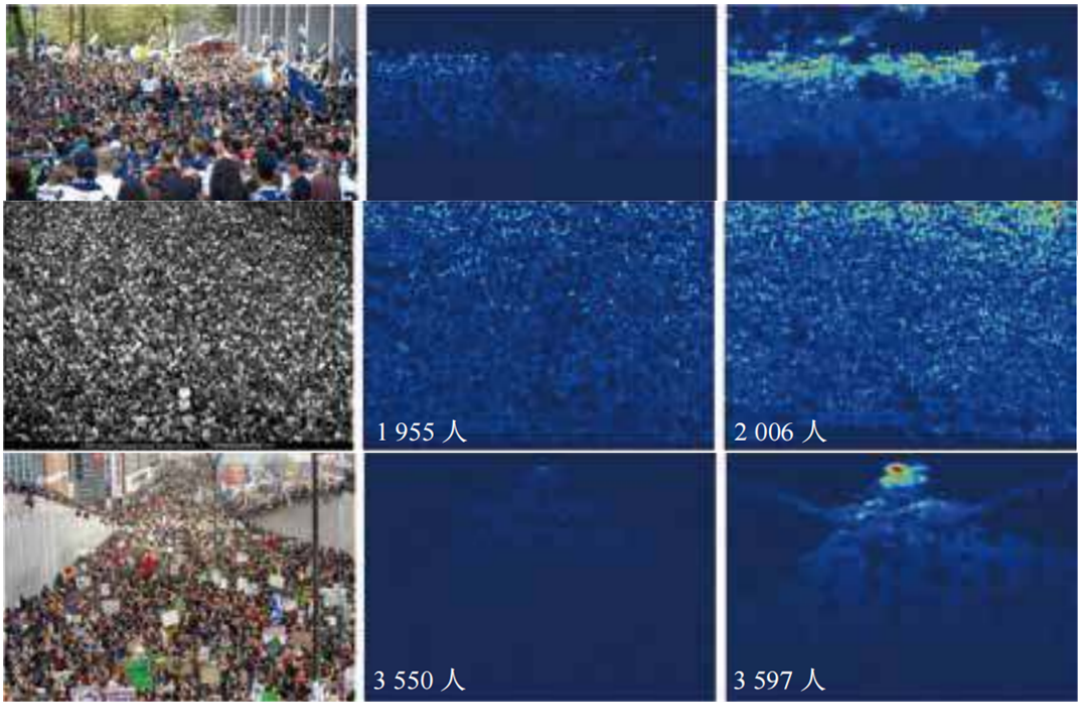

Fig. 3. Partial visualization results

3.6. Ablation Experiment

To further verify that CAM can effectively improve the model performance, ablation experiments were conducted on the original PPM and CAM structures on the Shanghai Tech dataset Part-A, and the results are shown in Table 6.

Table 6. Ablation experiments of CAM structure

Method	MAE	RMSE
Proposed+PPM	63.7	105.5
Proposed+CAM	62.6	95.8

As can be seen from Table 6, the performance has been improved when the CAM structure is used in this paper, with MAE reduces by 1.1 and RMSE reduces by 9.7. Because CAM can effectively promote the interaction of feature information at various scales, enhance context awareness, and improve the robustness of the model.

The proposed model is mainly composed of three modules: CAM, DCM and MSAM. To further verify the rationality and effectiveness of the structure of each Part, CAM, DCM, CAM+MSAM, DCM+MSAM, DCM+CAM

Table 7. Ablation experiments of different module structures

Method	MAE	RMSE
CAM	68.3	118.9
DCM	66.3	113.1
DCM+CAM	65.0	109.9
CAM+MSAM	65.6	111.5
DCM+MSAM	64.1	111.6
DCM+CAM+MSAM	62.6	95.8

and DCM+CAM+MSAM are respectively tested on the ShangHai Tech dataset Part-A. Ablation experiment results are shown in Table 7.

It is not difficult to find from Table 7 that both a single CAM and a DCM can obtain population information to a certain extent, but their counting accuracy is relatively low. After fusing DCM and CAM, the counting performance has been improved, indicating that the multi-scale information extracted by the fused structure is richer. The proposed method, based on the fusion structure, densely connects four layers of FEB structures to enhance the information transmission of each network layer and improve the modeling ability of continuous scale changes, achieving the best counting results in both MAE and RMSE. In addition, after CAM and DCM are combined with MSAM respectively, MAE decreases to 65.6 and 64.1 respectively, and RMSE decreases to 111.5 and 111.6 respectively, indicating that the MSAM structure has a good enhancing effect on feature aggregation.

4. Conclusion

A dense crowd counting network based on multi-scale perception is proposed and its performance is verified on four benchmark datasets. Its evaluation index is superior to other comparison methods, and it has good counting accuracy and robustness for different dense crowd images. The multi-scale feature aggregation module cascades four feature enhancement blocks in a dense residual connection manner, effectively aggregating cross-level features and improving the continuity of multi-scale information. The ablation experiment has fully demonstrated that this module can effectively enhance the modeling ability of continuous scale changes. The context-aware module designed based on the pyramid pooling structure can promote the interaction of multi-scale information among various branches and enhance the expression of multi-scale context information. In subsequent work, from the perspective of enhancing the robustness of the algorithm to background information, an attention mechanism will be introduced to better focus on crowd areas, further weakening the influence brought by background interference and chaotic crowd distribution, and improving the performance of crowd counting.

5. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. None.

References

1. Zhou W, Yang X, Dong X, et al. MJPNet-S*: Multistyle joint-perception network with knowledge distillation for drone RGB-thermal crowd density estimation in smart cities[J]. IEEE Internet of Things Journal, 2024.
2. Zhou W, Yang X, Yan W, et al. Hybrid knowledge distillation for RGB-T crowd density estimation in smart surveillance systems[J]. IEEE Internet of Things Journal, 2024.
3. Wang M, Zhou X, Chen Y. A comprehensive survey of crowd density estimation and counting[J]. IET Image Processing, 2025, 19(1): e13328.
4. Mei L, Yu M, Jia L, et al. Crowd Density Estimation via Global Crowd Collectiveness Metric[J]. Drones, 2024, 8(11): 616.
5. Lin C, Hu X. Efficient crowd density estimation with edge intelligence via structural reparameterization and knowledge transfer[J]. Applied Soft Computing, 2024, 154: 111366.
6. Li Y C, Jia R S, Hu Y X, et al. A weakly-supervised crowd density estimation method based on two-stage linear feature calibration[J]. IEEE/CAA Journal of Automatica Sinica, 2024, 11(4): 965-981.
7. Li Y C, Jia R S, Hu Y X, et al. A lightweight dense crowd density estimation network for efficient compression models[J]. Expert Systems with Applications, 2024, 238: 122069.

8. Ranasinghe Y, Nair N G, Bandara W G C, et al. CrowdDiff: Multi-hypothesis crowd density estimation using diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 12809-12819.
9. Gupta I, Seeja K R. Crowd Density Estimation for Video Surveillance Using Deep Learning: A Review[C]//International Conference on Smart Computing and Communication. Singapore: Springer Nature Singapore, 2024: 293-305.
10. Seo J, Choi J, Lee J. Adaptive transit control with crowd density estimation for uwb contactless gate[C]//2024 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2024: 878-883.
11. Wang S, Lyu Y, Li Y, et al. MIANet: Bridging the Gap in Crowd Density Estimation With Thermal and RGB Interaction[J]. IEEE Transactions on Intelligent Transportation Systems, 2024.
12. Alotaibi S R, Mengash H A, Maray M, et al. Integrating Explainable Artificial Intelligence with Advanced Deep Learning Model for Crowd Density Estimation in Real-world Surveillance Systems[J]. IEEE Access, 2025.
13. Patidar M, Bhanodia P K, Patidar P K, et al. Advanced Crowd Density Estimation Using Hybrid CNN Models for Real-Time Public Safety Applications[J]. Library of Progress-Library Science, Information Technology & Computer, 2024, 44(3).
14. Jaiswal S, Gadgil A S, Kaslikar A M, et al. Comprehensive Study of Various Methods for Estimating Crowd Density[C]//International Conference on Innovations and Advances in Cognitive Systems. Cham: Springer Nature Switzerland, 2024: 383-400.
15. Kamra V, Vaishnav A, Verma A, et al. A Novel Approach for Crowd Analysis and Density Estimation by Using Machine Learning Techniques[C]//2024 International Conference on Intelligent Systems for Cybersecurity (ISCS). IEEE, 2024: 1-6.
16. Tahira N J, Suresh P D, Park J S. Deep Learning based Approach for Crowd Density Estimation and Flow Prediction[C]//2024 24th International Conference on Control, Automation and Systems (ICCAS). IEEE, 2024: 1274-1275.
17. Tripathy S K, Srivastava S, Bajaj D, et al. A Novel cascaded deep architecture with weak-supervision for video crowd counting and density estimation[J]. Soft Computing, 2024, 28(13): 8319-8335.
18. Al Hossain F, Tonmoy M T H, Lover A, et al. Crowdotic: A Privacy-Preserving Hospital Waiting Room Crowd Density Estimation with Non-speech Audio[C]//Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications. 2024: 79-85.
19. Hu Y, Lin Y, Yang H, et al. CLDE-Net: crowd localization and density estimation based on CNN and transformer network[J]. Multimedia Systems, 2024, 30(3): 120.
20. Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
21. Zhang R, Chen J, Feng L, et al. A refined pyramid scene parsing network for polarimetric SAR image semantic segmentation in agricultural areas[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.
22. Elharrouss O, Mohammed H H, Al-Maadeed S, et al. Crowd density estimation with a block-based density map generation[C]//2024 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE, 2024: 1-7.
23. Trung H D. Estimation of Crowd Density Using Image Processing Techniques with Background Pixel Model and Visual Geometry Group[J]. Buletin Ilmiah Sarjana Teknik Elektro, 2024, 6(2): 142-154.
24. Yi C, Cho J. Robust Estimation of Crowd Density Using Vision Transformers[J]. International Journal on Advanced Science, Engineering & Information Technology, 2024, 14(5).
25. Bhatt C, Kukreti A, Pratap A, et al. Deep Learning for Crowd Counting: Addressing Crowd Density with Advanced Methods[C]//2024 Second International Conference on Advances in Information Technology (ICAIT). IEEE, 2024, 1: 1-5.
26. Khushi K, Jagriti S. Crowd Density Estimation Using Deep Learning: A Convolutional Neural Network Approach for Real-time Monitoring[J]. International Journal of Trend in Scientific Research and Development, 2024, 8(5): 472-476.
27. Alhawsawi A N, Khan S D, Ur Rehman F. Crowd counting in diverse environments using a deep routing mechanism informed by crowd density levels[J]. Information, 2024, 15(5): 275.
28. Jayasingh S K, Naik P, Swain S, et al. Integrated crowd counting system utilizing IoT sensors, OpenCV and YOLO models for accurate people density estimation in real-time environments[C]//2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU). IEEE, 2024: 1-6.
29. Sharma V K, Mir R N, Singh C. Scale-aware CNN for crowd density estimation and crowd behavior analysis[J]. Computers and Electrical Engineering, 2023, 106: 108569.
30. Tatlcan D, Apaydin N N, Yaman O, et al. Crowd density estimation via a VGG-16-based CSRNet model[J]. Inf. Dyn. Appl, 2025, 4(2): 66-75.
31. Wang W, Liu Q, Wang W. Pyramid-dilated deep convolutional neural network for crowd counting[J]. Applied Intelligence, 2022, 52(2): 1825-1837.
32. Khan M A, Menouar H, Hamila R. LCDnet: a lightweight crowd density estimation model for real-time video surveillance[J]. Journal of Real-Time Image Processing, 2023, 20(2): 29.
33. Wan J, Wang Q, Chan A B. Kernel-based density map generation for dense object counting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(3): 1357-1370.
34. Lian D, Li J, Zheng J, et al. Density map regression guided detection network for rgb-d crowd counting and localization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1821-1830.
35. Liu J, Gao C, Meng D, et al. Decidenet: Counting varying density crowds through attention guided detection and density estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5197-5206.
36. Jian C, Lin C, Hu X, et al. Selective Scale-Aware Network for Traffic Density Estimation and Congestion Detection in ITS[J]. Sensors, 2025, 25(3): 766.

37. Liang L, Zhao H, Zhou F, et al. SC2Net: scale-aware crowd counting network with pyramid dilated convolution[J]. Applied Intelligence, 2023, 53(5): 5146-5159.